

Seasonality of Circulation in Southern Africa using the Kohonen Self-Organising Map

by

Jeremy PL Main

Submitted in partial fulfilment
of the requirement for
the degree of

Master of Science

in the
Department of Environmental and
Geographical Sciences at the
University of Cape Town,
South Africa

May 1997

The University of Cape Town has been given
the right to reproduce this thesis in whole
or in part. Copyright is held by the author.

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

Abstract

A technique employing the classification capabilities of the Kohonen self-organising map (SOM) is introduced into the body of computer-based techniques available to synoptic climatology. The SOM is one of many types of artificial neural networks (ANN) and is capable of unsupervised learning or non-linear classification.

Components of the SOM are introduced and an application is then illustrated using observed daily sea level pressure (SLP) from the Australian Southern Hemisphere data set. To put the technique in the context of global climate change studies, a further example using simulated SLP from the GENESIS version 1.02 General Circulation Model (GCM) is illustrated, with the emphasis on the ability of the technique to highlight differences in seasonality between data sets.

The SOM is found to be a robust technique for deducing the modes of variability of map patterns within a circulation data set, allowing variability to be expressed in terms of inter- and intra-annual variability. The SOM is also found to be useful for comparing circulation data sets and finds particular application in the context of global climate change studies.

Acknowledgements

I would like to acknowledge; The Water Research Commission (WRC) of South Africa for funding support during the period of this study; Dr Bruce Hewitson, my thesis supervisor, for his continued enthusiastic and genuine support and; Prof Crane of The Pennsylvania State University for his insightful discussions and comments. For contributions of a more direct nature, I would like to acknowledge Zoë Christodoulou, Ann McMaster, my Parents and my brother for all the support in matters of the mind and soul. Finally, I would like to acknowledge my colleagues Debbie Hudson and Debbie Shannon and associates Kevin Winter, Peter Holmes and Margot Jansen van Nieuwenhuizen for all their unwavering assistance.

Table of Contents

ABSTRACT.....i

ACKNOWLEDGEMENTS.....ii

TABLE OF CONTENTS iii

LIST OF FIGURES v

LIST OF TABLESvii

CHAPTER 1: INTRODUCTION 1

1.0 INTRODUCTION..... 1

1.1 SYNOPTIC CLIMATOLOGY.....4

1.2 AIMS AND APPROACH7

CHAPTER 2: SYNOPTIC CLIMATOLOGY AND SELF-ORGANISING MAPS.....11

2.0 SYNOPTIC CLIMATOLOGY..... 11

2.1 OVERVIEW OF THE KOHONEN SELF-ORGANISING MAP 16

 2.1.0 *Introduction to Artificial Neural Networks*..... 16

 2.1.1 *Kohonen's Self-Organising Map* 22

2.2 GCM SIMULATIONS 26

2.3 SUMMARY OF PROCEDURES 30

CHAPTER 3: DATA PREPARATION AND METHODS.....32

3.0 STUDY AREA AND DATA PREPARATION 32

3.1 METHODS AND PROCEDURES OF THE SOM 37

 3.1.0 *Structural Considerations*..... 37

 3.1.1 *Training Considerations*..... 39

 3.1.2 *Training*..... 43

 3.1.2.0 Random Initialisation 43

 3.1.2.1 Linear Initialisation 44

3.1.3 *Post-Training Analysis and Visualisation* 44

3.1.3.0 Decomposing Weights.....45

3.1.3.1 Decomposing Nodes.....45

CHAPTER 4: APPLICATION OF THE SOM46

4.0 APPLICATION OF THE SOM.....46

4.1 OBSERVED CIRCULATION47

4.2 DOUBLED CARBON DIOXIDE CIRCULATION.....62

4.3 DISCUSSION AND SUMMARY72

CHAPTER 5: CONCLUSION73

5.0 DISCUSSION.....73

5.1 CONCLUSION.....74

REFERENCES77

List of figures

Figure 2.0 : The general structure of an artificial neural network (after Hewitson and Crane 1994).	
Each node in the network is interconnected (omitted in diagram for clarity).....	18
Figure 2.1 : The general structure of a competitive or unsupervised ANN. The structure is	
characterised by a single output layer arranged as a two-dimensional array of nodes, each	
connected with an in-line bias or weight to each of the inputs. In this example vector x is one	
sample of the input vectors that make up a data set.	21
Figure 2.2 : The SOM using the optical lens analogy. Input vectors are projected onto the two-	
dimension node map (Kohonen map).....	24
Figure 2.3 : Mapping a circulation pattern to node weights in a trained SOM. An input gridded	
circulation pattern is presented to the SOM (a), the nodes compete to win the pattern by having	
the best match (b), the winning node calculates an error (c) and the location of the associated	
node on the two-dimensional map is recorded (d).....	25
Figure 3.0 : Map of study area (inner box) and layout of the GENESIS R15 grid points (circles) and	
the Australian Southern Hemisphere observed gridded data set (squares).....	34
Figure 3.1 : Two possible topologies a) hexagonal and b) rectangular, that may be selected for the two-	
dimensional node map or SOM.	38
Figure 3.2 : a) The Gaussian neighbourhood kernel updates weights within a radius of the winning	
node except for the update rate that increases toward the centre of the radius, approximating a	
'bell curve', this figure shows three variations of the structure determined by sigma. b) The	
bubble neighbourhood kernel evaluates the weights only if they are in the update neighbourhood,	
weights outside that area are left intact.	40
Figure 3.3 : Graph showing the way in which the learning rate (α) is reduced with the linear and	
inverse_ t schemes over 1000 iterations.	41
Figure 4.0 : Sammon mapping from the second training with Observed SLP circulation data.	48
Figure 4.1 : Sammon mapping from the ninth training with Observed SLP circulation data.	49
Figure 4.2 : Time series of errors calculated from the snapshots of the weights during training number	
2.	50

Figure 4.3 : Meta-map for Observed SLP circulation.....	51
Figure 4.4 : A sample of the best matching input circulation pattern from June 1979 won by node (4,6)	53
Figure 4.5 : A sample of the best matching input circulation pattern from January 1974 won by node (1,1)	54
Figure 4.6 : Monthly frequencies and error map for Observed SLP circulation patterns.	56
Figure 4.7 : Trajectory plots of centroids for each month from observed SLP.	58
Figure 4.8 : Monthly frequencies and error map for control simulation SLP circulation patterns.....	59
Figure 4.9 : Trajectory plots of centroids for each month from GENESIS GCM control SLP.....	60
Figure 4.10 : Time series of errors calculated from the snapshots of the weights during training for the sixth run.	62
Figure 4.11 : Sammon mapping from the sixth training with GENESIS GCM control run simulation circulation data.	63
Figure 4.12 : Meta-map for GENESIS GCM control run SLP circulation.....	65
Figure 4.13 : Monthly frequencies and error map for control simulation SLP circulation patterns.....	66
Figure 4.14 : Trajectory plots of centroids for each month of GENESIS GCM control SLP.	67
Figure 4.15 : Monthly frequencies and error map for doubled CO₂ simulation circulation patterns. ...	69
Figure 4.16 : Trajectory plots of centroids for each month of GENESIS GCM doubled CO₂ SLP.....	70

List of tables

Table 3.0 : Summary statistics of the three circulation (SLP) data sets before standardisation. Mean and standard deviation are calculated with respect to the entire data set (matrix-wise).37

Table 4.0 : Observed SLP Frequency map statistics.57

Table 4.1 : GCM control run SLP frequency map statistics.....60

Table 4.2 : GCM control run SLP frequency map statistics.....64

Table 4.3 : GCM doubled CO₂ run SLP frequency map statistics.68

CHAPTER ONE

Introduction

1.0 Introduction

Southern Africa may be defined as the area south of the equatorial region of Africa, and except for the extreme south-west, which receives winter rainfall, the region is a summer rainfall region. The climate is characterised by relatively low amounts of annual rainfall in the predominantly summer rainfall regions, with a high degree of inter-annual and intra-annual variability, which aggravates the reliability of rainfall. This has significant socio-economic implications due to the increasing demand for water from a growing population and the considerable pressure placed on water resources. Agriculture is perhaps most vulnerable since, for example, about 50% of South African water resources are consumed by irrigation¹, while those crops which are not irrigated are dependent upon the seasonality of the rainfall, for their success or failure. Other climatic factors such as surface humidity, soil moisture and temperatures, naturally affect aridity and, like rainfall, fluctuate on inter-annual and intra-annual time scales.

Variability within the climate system is a natural phenomenon and is socially and environmentally significant and continuous on all time scales. However, during this century there has been increasing concern about the ability of humans to induce significant

¹ See URL <http://www.anc.org.za/water/conserve.html> for South African Water usage statistics.

change in the climate system by their activities. In particular, the Intergovernmental Panel on Climate Change (IPCC) has reported on recent increases in mean global carbon dioxide levels, which have been linked directly to human activities. The pre-industrial (1750 AD) concentration value of 280 ppmv has increased to the present day concentration value of 360 ppmv (Houghton *et al.*, 1992²). Carbon dioxide is classified as a greenhouse gas, due to its radiation modifying properties, and in the climate system plays a role of retaining energy in the atmosphere that would otherwise be radiated out to space. In theory, increased levels of carbon dioxide are expected to increase global temperatures, however, due to the interrelationships that exist in the climate system, the very nature of the global climate dynamics would also be expected to change. Just how that change will manifest itself in terms of surface characteristics is an issue of much debate, and particularly how such changes may affect regional climates in southern Africa.

Numerous studies have been undertaken focusing on the variability of the circulation over southern Africa and the resulting effects on the climate (e.g. Taljaard, 1986; Tyson 1984, 1981; Theron and Harrison, 1991). Similarly, there have been many studies focusing on boundary layer processes and causal mechanisms and in particular rainfall characteristics at the surface (e.g. Hulme, 1992; Nicholson, 1986; Matarira and Jury, 1992; Matarira, 1990). Southern African rainfall has been shown to have a strong dependence on surrounding sea surface temperatures (SSTs) as well as with SSTs in the Pacific Ocean through teleconnection mechanisms (Mason and Lindesay, 1993; Mason and Tyson, 1992; Walker,

² The 1995 executive summary is currently available on the internet at URL <http://www.unep.ch/ipcc/>

1990). It has also been shown that the El Niño Southern Oscillation (ENSO) impacts with a degree of spatial and intra-annual variability upon the summer rainfall region of southern Africa (D'Arbreton and Lindesay, 1993; Lindesay, 1988). Furthermore, key cycles have been associated with rainfall variability over the region for example, ENSO (Lindesay and Vogel, 1990), the quasi-biennial oscillation (QBO) (Brankovic *et al.*, 1994; Mason and Lindesay, 1993), and the semi-annual oscillation (SAO) (Hurrell and van Loon, 1994), as well as other lower frequency forcings such as the 18-year cycle (Mason, 1990) and 10-11 year cycles (Currie, 1993). Recent work using signal processing techniques show evidence of changes in both the Northern Hemisphere average temperature and in the seasonal cycle. For example, the seasons in central England have been delayed by 4.5 days in the last 50 years, which is as much as they have been delayed in the last 300 years (Kerr, 1995).

Probable global climate change related to anthropogenic influences could further aggravate the reliability of rainfall and is thus one of the primary focuses of southern African climate change research. Variability on the annual and intra-annual time scales has implications for human economic activities and hence the research interest in mechanisms and behaviours that affect the timing, onset and magnitude of the seasons (e.g. Jolliffe, 1994; Bayo Omotosho, 1992). Understanding the broad groups of atmospheric conditions and how they manifest on the surface, with particular reference to their seasonality and frequency, could allude to the processes taking place within the climate system that cause them. Clearly, a greater understanding of these processes would be of considerable value to policy makers and planners by providing a framework within which to react to global climate change.

Within the context of the issues of global climate change, vulnerability and its possible impacts, this study sets out to examine how these issues may have been approached in the past using the methods of synoptic climatology, paying some attention to a review of methods and their limitations. A new method is then introduced using the Kohonen Self-Organizing Map. This method is applied to a global climate change question in which climate model simulated data are validated against observed data and climate change scenarios are subsequently derived for the southern African region.

1.1 Synoptic Climatology

The task of simplifying the complex variability of the weather into general modes of behaviour has been the focus of climatology for some time and, in particular, the relation of modes of behaviour of the weather to conditions at the Earth's surface. This has been termed synoptic climatology and addresses questions that are concerned with relating general characteristics in the atmosphere to elements of the surface environment. More recently, these kinds of questions have included those which have arisen out of climate change studies.

Synoptic climatology is the result of the merging of a number of fields of climatology incorporating dynamic climatology, climatic change, regional climatology, physical climatology and applied climatology (Yarnal, 1993). The result is a field of climatology that uses regional scale studies of climate in conjunction with the interpretation of associated physical processes, in an applied sense, to answer questions about the relationship of synoptic scale circulation patterns to regional scale surface environments.

The applied nature of synoptic climatology and the regional scale of study has consequently made the field popular with geographers.

Classification of general weather conditions into common types is fundamental to all approaches in synoptic climatology. These techniques have included manual classification of synoptic charts on the basis of map patterns and subjective selection of parameters that group surface stations together. Many of these earlier techniques were formalised in *Synoptic Climatology: Methods and Applications*, authored by Barry and Perry in 1973. A recent successor to Barry and Perry's work on methods of synoptic climatology is entitled *Synoptic Climatology in Environmental Analysis: A Primer* by Yarnal in 1993. In the 20 years between these two books, large shifts have taken place between the way in which synoptic climatology was initially tackled and the wide collection of computer assisted techniques subsequently introduced. Computing power has made a significant impact on synoptic climatology, and methods presented in Barry and Perry's book have been extended to incorporate automated procedures and digital environmental data. Yarnal (1993) describes automated procedures as those which may also be referred to as computer based or computer assisted methods. A common thread, however, in both manual and 'objective' computer based synoptic climatology techniques is the subjectivity used to define parameters and interpret results (Key and Crane, 1986).

Manual classification, on the one hand, requires the investigator to subjectively define categories by visual inspection of the map patterns. In computer based classifications, on the other hand, subjective decisions are made such as defining how many categories the

investigator would retain in a Cluster Analysis. Thus, no matter how objective a technique is, there will always be one point in a procedure where the investigator will need to make a subjective decision that may have serious implications for the final results of the analysis.

The discrete nature in which groups within the classification are sharply defined do not allow multi-group membership and thus impose artificial boundaries on what is, in essence, a continuum of categories. Consequently, the tendency of synoptic climatological techniques has been to generalise, with the result that extreme events are lumped into more general categories. In addition, some techniques, for example, eigenvector techniques, while mathematically compact and 'objective', impose their own particular problems in interpretation.

Nonetheless, synoptic climatology has found wide application, although it has not always been the optimal approach for the climate problems to which it has been applied. By their very nature, techniques in synoptic climatology are designed to move from the specific to the general, taking complex data and producing generalised summaries of the properties and characteristics of that data. It is, therefore, not a particularly effective approach for the analysis of specific events, but should rather be used for long-term analysis of climate states, such as in climate change studies.

To address some of the shortcomings of more traditional techniques, the following work introduces an alternative classification methodology using the Kohonen Self-Organising Map (SOM). The Self-Organizing Map Program Package (SOM_PAK) was developed by

Teuvo Kohonen of the Helsinki University of Technology's Laboratory of Computer and Information Science, in Finland and focuses on pattern recognition, yet has not been applied in synoptic climatology. A worked example, using southern African circulation data and General Circulation Model (GCM) data is conducted in this thesis to illustrate the procedures of the SOM and explore some of the possibilities of this technique in the context of elevated atmospheric carbon dioxide (CO₂) and global climate change studies.

South Africa provides a good example for testing a new synoptic climatology procedure given the highly variable nature of the regional climate system. Furthermore, southern Africa is a particularly important region for analysis due to the high degree of societal vulnerability to global change.

1.2 Aims and Approach

The Southern Hemisphere responds differently from the Northern Hemisphere to climate forcing factors due to its unique composition of land, sea and ice, of which land is the least significant. Unfortunately, most of the research involving GCMs has been carried out in the Northern Hemisphere resulting in models that have been fine-tuned for the Northern Hemisphere, while less attention has been paid to performance in the Southern Hemisphere (Whetton *et al.*, 1996).

GCMs need to be validated against observed data, and in the southern hemisphere this is complicated by the fact that, due to the scarcity of land, observed data sets are not of

comparable quality to Northern Hemisphere data. Comparing and making statements about the differences between the observed and simulated environments is problematic and is the subject of considerable research (Gates, 1992). Generally, inter-model comparisons are made in terms of differences in the variability about the long-term means. This is, at best, a limited validation.

In general, GCM based climate change studies have shown that global climate change is as likely to manifest itself in changes in variability as it is in overall changes in means and, more importantly, may be very regionally specific (e.g. Houghton *et al.*, 1992; Hewitson and Crane, 1996; Hewitson and Main, 1996). In particular, attention needs to be paid to scenario development at the regional scale, with emphasis on event frequency, persistence and variability, as well as changes in means (e.g. Tyson and Hewitson, 1996). While techniques other than GCMs exist for the development of regional scenarios (e.g. palaeo-analog studies), climate system modelling with GCMs are likely to prove the most productive in terms of developing regional scale climate change scenarios (Joubert and Hewitson, 1996; Tyson, 1993).

However, due largely to the intensive computational requirements, GCMs are normally run with a coarse spatial resolution in the order of 2.75 to 3.5 degrees of latitude and longitude respectively. Regional (sub-grid) scale features of climate are thus not explicitly represented. Additional techniques are thus required to obtain sub-grid-scale information, such as empirical downscaling (e.g. Hewitson and Crane, 1996; Hewitson, 1994; von Storch, 1993; Hewitson and Crane, 1992a; Hewitson and Crane, 1992b; Grotch and

MacCracken, 1991; Wigley *et al.*, 1990) or nested modelling (e.g. Mearns *et al.*, 1995a; Mearns *et al.*, 1995b; Giorgi *et al.*, 1994). However, these depend upon synoptic scale data from the GCM, and an approach that directly evaluates the synoptic characteristics would greatly assist in the application of the GCM data.

In this context the aim of the current study is to introduce a methodology using the Kohonen Self-Organising map (SOM), which assists in circumventing some of the difficulties of the more traditional synoptic climatological techniques. In so doing, this study will investigate the inter- and intra-annual variability of the atmospheric circulation over Southern Africa in the recent past and in GCM simulations of present and future climates. Thus, using the Kohonen SOM, a synoptic climatology analysis of observed and GCM data is performed in order to investigate GCM performance at the scale at which GCMs perform best.

Furthermore, while the inter-/intra-annual variability of the climate system in the past is observable and available for study in the historical and reconstructed records, direct observation of future trends is not possible. The best hope of establishing the likely impact of global climate change is left to the General Circulation Model (GCM). In this regard, applied synoptic climatology techniques are particularly pertinent to South Africa as a tool for establishing the relationship between the country's generally arid climate and the general circulation, and for investigating the characteristics of synoptic scale circulation.

As regional climates are largely a function of synoptic scale forcing, analysis of the synoptic scale in GCMs offers valuable insight into possible changes in regional climate forcing. Hence, synoptic climatology is a useful interpretation tool for establishing the characteristics of particular synoptic classes and their variability between and within years. A short review of synoptic climatology methods is initially presented in order to establish the framework within which the SOM methodology may be introduced. The SOM is then described and the general procedure of the SOM is shown with reference to global climate change studies and the use of the GCM. The remainder of the study will illustrate the use of the SOM with a specific example involving southern African and simulated GCM circulation data, in which seasonal circulation signals are expressed in terms of fluctuations across classification categories of the SOM and data sets are thus compared.

CHAPTER TWO

Synoptic Climatology and Self-Organising Maps

2.0 Synoptic Climatology

Synoptic climatology is a sub-discipline of climatology which focuses on relating the large scale atmospheric circulation to the climate that is experienced at a place or region, and where the classification of circulation is a means by which generalisations of the weather may be made. Synoptic climatology has experienced a shift from manual to computer-based techniques in recent years, although this has not affected the primary definition and aims.

Climatology, on the broader scale, may be characterised by the study of long-term relationships in the weather. For example, Eagleman defines climatology as "*the study of the long-term atmospheric environment at the earth-atmosphere interface*" (Eagleman, 1976, p1). While climatology relates to generalisation of the weather through time, synoptic climatology relates to the system of methodologies and approaches for determining the nature of a climatology. Barry and Perry (1973) note that the term synoptic climatology was first proposed in 1942 and point out that humans have practised these methods since agriculture played an important role in their lives. In summarising the aim of synoptic climatology, Barry and Perry (1973) state that "*[Synoptic climatology] is to relate local or regional climates to a meaningful frame of reference - the atmospheric circulation...*" (Barry and Perry, 1973, p7). More specifically "*the field of synoptic*

climatology is concerned with obtaining insight into local or regional climates by examining the relationship of weather elements, individually or collectively, to atmospheric circulation processes." (Barry and Perry, 1973, p5).

Yarnal (1993) proposes the following working definition: "*synoptic climatology relates the atmospheric circulation to the surface environment*" (Yarnal, 1993, p 5), and continues by adding that this field is "*important to other atmospheric sciences because it synthesizes several fields of climatology*" (Yarnal, 1993, p 1). The surface environment is inclusive of the planetary boundary layer, within which large scale atmospheric motions are assumed to manifest themselves. In this context, the relationships are identified through classifying data into a few discrete types in order to aid interpretation. The atmosphere, however, is a multi-dimensional continuum, which presents a problem for discrete classification, and may result in over simplification of map patterns.

The general aims of synoptic climatology are related to classification and cross-scale analysis and in so doing establish the effects of climate variability on the surface environment. Thus, there are two approaches to synoptic classification as proposed by Yarnal (1993): circulation-to-environment and environment-to-circulation, each one determining whether the circulation data or the environmental data are classified respectively. In the first approach, once the circulation data are classified, the synoptic classes are related to the environmental data. Alternatively, the environmental data may be classified and may become the determinant for synoptic classes (Yarnal 1993). In both approaches, classification is a common procedure in which techniques vary and implementation is either manual or computerised.

Manual classification refers to all classifications that are not automated and represent methodologies used at the inception of synoptic climatology. Manual classification is labour intensive and it is unlikely that two investigators will reach the same results as the basis for definition of synoptic groups are determined subjectively. Early manual techniques entailed direct manual classification of daily weather charts into distinct groups, and those that were less easily discernible were assigned to hybrid or unclassifiable groups. For example, Lamb (1972) devised a daily weather-map classification, comprising seven types, for use over the British Isles. Lamb's study extended from 1861 to 1971 and is internally consistent since the results were derived by one investigator. Similarly, Muller (1977) devised a generic manual synoptic-type classification explicitly to relate the surface environment and human activities around the United States Gulf Coast. The Muller types extend from 1951 to the present and are characterised by synoptic types that show seasonal cycles, inter-annual variability and geographic variation of these types within the region.

Computer-based techniques are characterised, to some extent, by automation, however, the decisions that are made by the investigator render the procedures subjective, as with manual techniques. Computer-based techniques typically manipulate pressure map patterns in a numeric digital form, of which correlation, eigenvector and cluster based procedures are most common. Lund (1963) and Kirchhofer (1973) are credited with the development of correlation-based techniques. Essentially, the technique computes a measure of similarity between two digital weather maps, Lund used the Pearson product-moment correlations and Kirchhofer used the sum-of-squares algorithm. The procedure is repeated for a large number of maps in order to derive a cross-correlation table which is then used to classify the maps into groups representing different synoptic types.

More recent progress in classification uses clustering in conjunction with eigenvector-based techniques. An example of such a classification is Principal Components Analysis (PCA). PCA is a mathematically compact technique for deriving orthogonal modes of variance (components) and significantly reduces the dimensions of the data. This can be used for both circulation-to-environment and environment-to-circulation studies.

Typically, the investigator computes eigenvalues and eigenvectors of the similarity matrix of a data set to determine its underlying structure. Following this, scores of derived components (eigenvectors) are clustered with a standard clustering procedure. Again, however, subjectivity plays an important role, for example in choosing the number of components, or choosing whether or not to rotate to avoid Buell patterns (Buell 1979), and often triggers much debate (e.g. Richman, 1986).

An alternative to clustering circulation data would be to perform an environment-to-circulation procedure. For example, Kalkstein and Corrigan (1986) devised a technique involving clustering the score time series resulting from a P-mode PCA of several variables of station data into groups that represent synoptic classes. The formulation is termed the Temporal Synoptic Index (TSI), from which clusters form the basis of an environment-to-circulation analysis, thus classifying circulation on the basis of air mass characteristics at one station. Kalkstein *et al.* (1987) later tested three clustering techniques, viz. Ward's minimum variance, average-linkage and centroid, of which the average-linkage was found to be the optimum method for such clustering.

The advent of computer-based synoptic climatology techniques was largely heralded as advantageous over manual techniques due to the objectivity of such techniques. This

misnomer was widely held until Key and Crane (1986) demonstrated the subjective manner in which correlation-based and eigenvector-based models were implemented. Key and Crane (1986) established that results of analyses were affected by decisions made by the investigator, for example the number of components retained in a PCA or the type of clustering method used. In conclusion, the authors proposed that since such methods were subjective, the investigator should be thoroughly familiar with the results as well as the data and climatology of the region in order to derive the best possible eigenvector solution.

Computing power has made a significant impact on synoptic climatology since the 1970's, and methods presented in Barry and Perry's 1973 book have subsequently been extended to incorporate automated procedures using computers. The further development of computer systems and their capabilities has resulted in exploration of other computer-based techniques including the use of artificial intelligence (AI) procedures, such as fuzzy sets for map pattern classification (Bardossy *et al.*, 1995) and artificial neural networks for establishing cross-scale relationships (Hewitson and Crane, 1996). It is into this group of computer-based procedures that the Kohonen self-organising map (SOM), described in the following section, is included.

Once the classification of circulation states, or the classification of surface climate variables has been completed, the synoptic types are typically related to specific conditions at the local or region scale, or to determine general circulation modes. It is in this area where the majority of synoptic climatology techniques rely on subjective decision making to derive categories, where there is seldom a continuous inter-relationship between categories, let alone between categories and surface climate. This is an aspect of synoptic

climatology that is addressed by the SOM, where categories are related to each other on a continuum across a two-dimensional cluster space. Aside from this useful feature, the SOM is also able to classify data on the basis of non-linear criteria.

2.1 Overview of the Kohonen Self-Organising Map

This study makes use of an implementation of the Self-Organising Map (SOM) as a classification procedure (Kohonen *et al.*, 1996; Kohonen, 1994, 1984)³. The SOM may be broadly described as an Artificial Neural Network (ANN) and has an inherent ability for classification which makes it potentially applicable for use in synoptic climatology.

Artificial Neural Networks (ANN) are currently being used in a number of research fields in Geography (Hewitson and Crane, 1994), and applications have ranged from prediction of precipitation and snowfalls from general circulation fields, to classification of Arctic cloud and sea-ice features from satellite data (Hewitson and Crane 1994).

2.1.0 Introduction to Artificial Neural Networks

The SOM is a special case of an ANN and, as such, the basic principles and terminology are common to all ANN structures. ANNs were initially intended as a mathematical representation of biological brain processes, although this is now seen to be the case only at the most superficial level. ANNs are capable of *learning* patterns and relationships, and

³ The particular software used in this study is available at the Internet site <http://nucleus.hut.fi/nnrc/nnrc-programs.html>. The most recent release at the time of writing is SOM_PAK Version 3.1 released on April 7, 1995. The accompanying documentation describes the theory and usage of the package.

can perform functions similar to a number of 'traditional' statistical functions, including what might be termed non-linear regression. An ANN is composed of an array of simple processors (nodes or units), each connected by communication channels (connections), which carry numeric data, and are biased by a weighting function (Clothiaux and Bachmann, 1994). Figure 2.0 illustrates a typical ANN, similar to those used in climate downscaling procedures (Hewitson and Crane, 1996).

As configured in figure 2.0, the ANN relates a multi-dimensional input to a one-dimensional output. Data is passed from nodes in the input layer through nodes of a *hidden* layer to the output layer. Each node is completely connected to every other node in adjacent layers, however, special configurations might implement only partial connectivity. The output function of each node may be varied, however it is normally a bounded non-linear function (e.g. a bipolar hyperbolic tangent), or it may be as simple as a linear or a threshold function.

In the case of a linear function, the ANN is functionally equivalent to linear multiple regression. In each node, the weighted summation of the respective node inputs is processed by the node function to generate a single output. The structural complexity of the neural network determines the degree to which a function may be represented (Wasserman, 1989).

As in simple linear regression, which is analogous in some sense to ANNs, a dependent variable (y) is related to an independent variable (x). Similarly, in the same way that more complex multiple regression procedures relate a set of dependent variables (y_1, \dots, y_m) to a

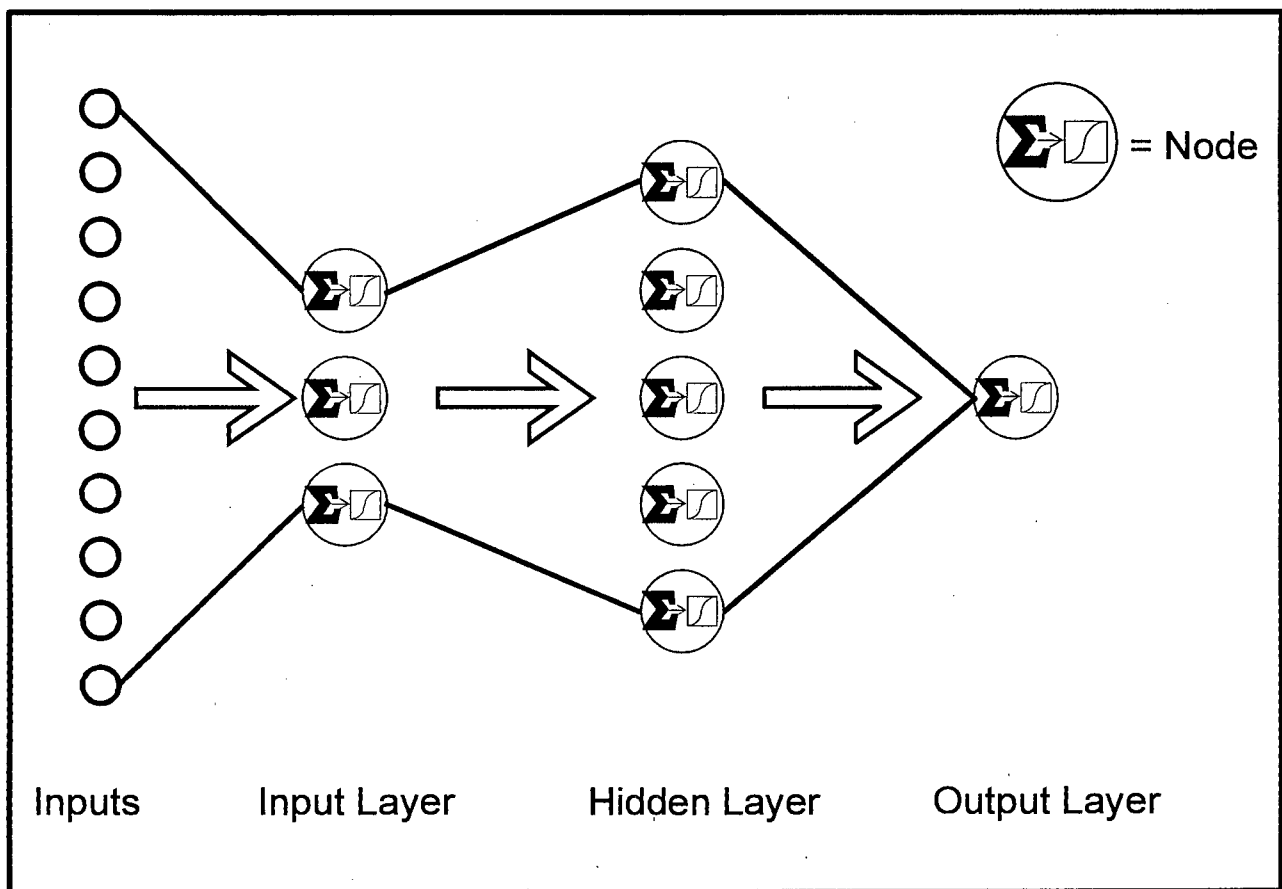


Figure 2.0 : The general structure of an artificial neural network (after Hewitson and Crane 1994). Each node in the network is interconnected (omitted in diagram for clarity).

set of independent variables ($x_1 \dots x_n$) via some linear or prescribed non-linear function, so are neural networks able to find a relationship between multiple variables and, importantly, without the constraint of linearity or prescribed non-linearity.

In practical terms, the artificial neural network expresses a function for mapping a set of inputs to set of outputs. The function between inputs, x , and outputs, y , is first derived by *training* an ANN. During training, a range of inputs are iteratively presented to the network where the connection weights are adapted according to some algorithm in an attempt to match the ANN output to some known output. The error between the predicted output and known output is used to monitor the training and provide a feedback to the network for adjustment of the weights. Training is complete when no improvement (reduction) can be made to the error. Algorithmically, this is simply a minimisation problem in which the error surface is minimised. The error surface is a function of the output of the ANN and known outputs for a given set of input data. A range of minimisation techniques are available from simple gradient descent to more complex approaches such as conjugate gradient techniques.

Once the ANN is trained, any set of input data (within the bounds of the training data set) may be presented to the net and the predicted values derived using the function represented by the ANN. The ability of the ANN to predict output values is restricted by the assumption that the input data lies within the bounds of the original training data.

However, unlike linear regression, the ANN has a particular advantage that, given suitable complexity, it may represent any linear, logical or non-linear function. Furthermore, ANNs have proved particularly robust in the presence of noise in the data and shown an ability to

generalise a function from limited training data (Hertz *et al.*, 1991). There are various implementations of ANNs, of which the supervised feed-forward ANN described above is one. A range of reference literature is available, and for further details see Hertz *et al.* (1991) or Wasserman (1989), or the ANN Frequently Asked Questions (FAQ) document on the Internet⁴.

In the absence of a target data set, a special case of the ANN, the competitive or unsupervised ANN, may be used. The task of these ANNs is to identify patterns in the input data with respect to the underlying structure within that data set. The unsupervised ANN comprises a specialised output layer of nodes which takes the form of a two-dimensional array (Figure 2.1). Each node in the network is connected via a weighted link to each of the inputs. Prior to training, the weights are initialised with starting values. One at a time the input vectors are then presented to the network and the values of all output nodes determined. The node with the least difference between itself and the input vector is termed the *winner* and becomes the centre of the *update neighbourhood*, the area within which nodes and their associated weights will be updated such that each weight vector converges to the input pattern (i.e. positive feedback). Within the update neighbourhood the degree of weight adjustment will decrease away from the winner node according to some spatio-temporal decay function or *kernel*. In this way the winner node tries to represent the particular input pattern, and surrounding nodes develop representations of similar, but not the same, patterns.

⁴ The latest version of the Frequently Asked Questions (FAQ) document, with ANN resources and starting points is available as a hypertext document under the URL <ftp://ftp.sas.com/pub/neural/FAQ.html> or in the news group URL <news://comp.ai.neural-nets>

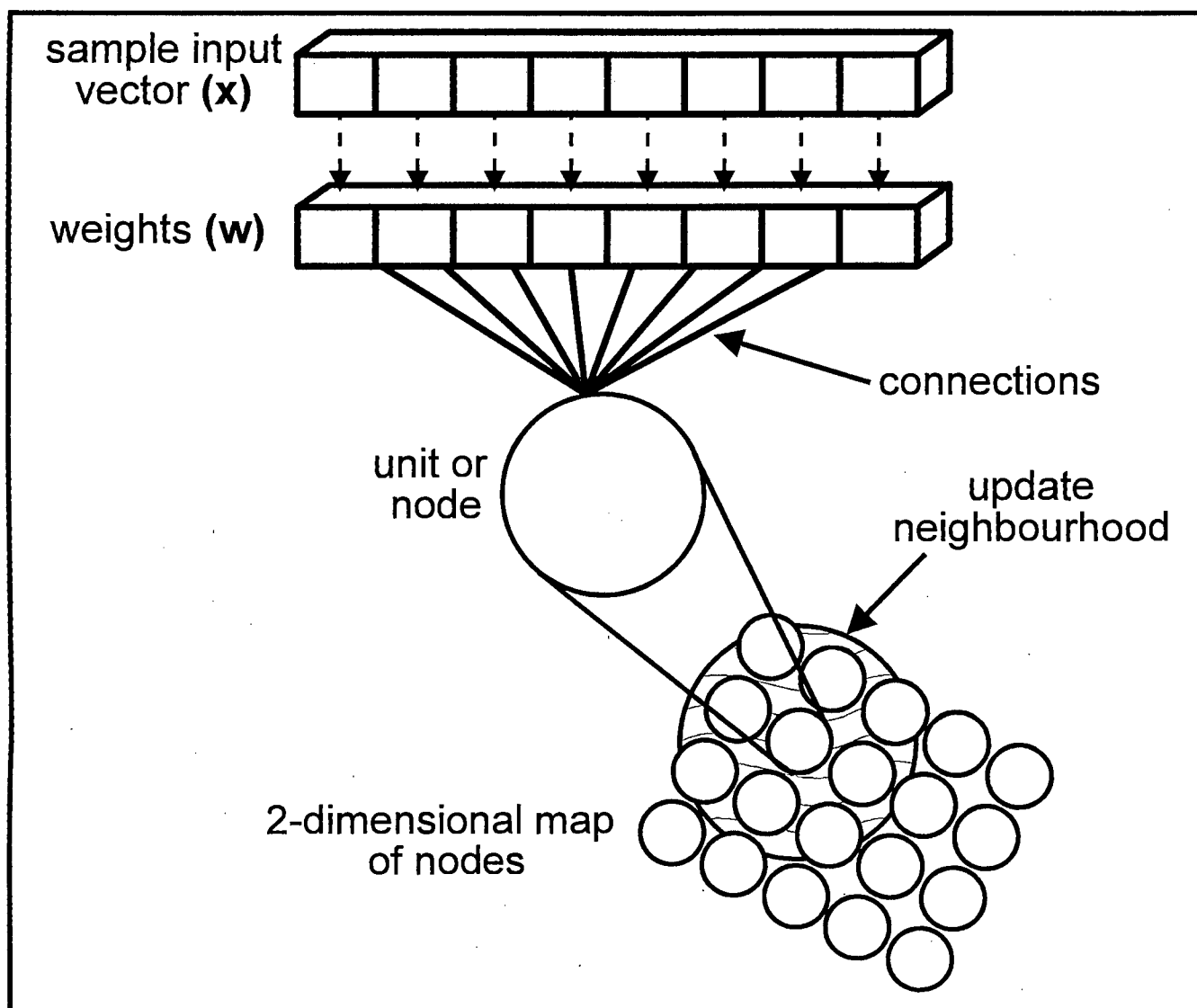


Figure 2.1 : The general structure of a competitive or unsupervised ANN. The structure is characterised by a single output layer arranged as a 2-dimensional array of nodes, each connected with an in-line bias or weight to each of the inputs. In this example vector \mathbf{x} is one sample of the input vectors that make up a data set.

The update neighbourhood of the kernel is thus responsible for bringing about a continuum across the two-dimensional map of nodes. The *kernel* regulates the rate at which the weights may be updated, i.e. the radius of the update neighbourhood, and the degree of weight adjustment as a function of distance from the winning node.

Once the unsupervised ANN is trained, the weights represent the general structure of the set of input vectors and take on values similar to those in the input data set. Samples of data, whose characteristics lie within the bounds of the training data set, may then be presented to the ANN, which maps each sample to a node in the two-dimensional node space with an associated measure of accuracy. The set of mappings, referred to as *visualisation* data, may then be interpreted with further analysis.

2.1.1 Kohonen's Self-Organising Map

The Kohonen Self-Organising Map (SOM) is an example of a competitive or unsupervised neural network and, as described above, uses no target data set for training. It is characterised by its ability to find the structure in the data itself (Kohonen, 1990). The SOM is distinguished from other ANNs by the inherent ability of cells within the network to compete with each other as shown in Figure 2.1. Thus, the SOM is a multiple winner competitive ANN with processing elements demonstrating a spatial structure.

One way to conceptualise the process represented by the SOM is to consider the analogy of an optical lens. In the same way that an optical lens re-maps a three dimensional image to a two-dimensional plane so does the SOM re-map high dimensional input vectors to a two-

dimensional node map (Figure 2.2). Similarly, as an optical lens projects nearby objects to spatially similar locations on the projection plane, so a SOM maintains some measure of the distance between data points in the high dimensional space. Properties of the SOM that are most pertinent to this project are the ability to a) classify input data based on non-linear relationships between the input elements and b) to express the categories in terms of a two-dimensional continuous space.

The main use for SOMs in the case of this study will be to characterise the spatial and temporal behaviour of sea level pressure (SLP) patterns. The SOM is trained with SLP circulation patterns to identify the circulation 'type' represented by each node (Figure 2.3). The particular input pattern represented by a trained node in the SOM array may be determined from the set of weights associated with each node. For each node there are as many weights as there are inputs and each grid point represents a point on a circulation map. In Figure 2.3, for example, a trained SOM is presented with a circulation map (Figure 2.3 a), and each node competes to try and best match the input circulation pattern (Figure 2.3 b). The winning node then calculates an error (Figure 2.3 c) which is recorded along with the location of the associated node on the two-dimensional node map (Figure 2.3 d).

While previous synoptic climatological methods imposed discontinuous categories on circulation patterns, the SOM node represents archetypes of input patterns spanning the continuum of the input range, where nodes close together represent related 'clusters' or types.

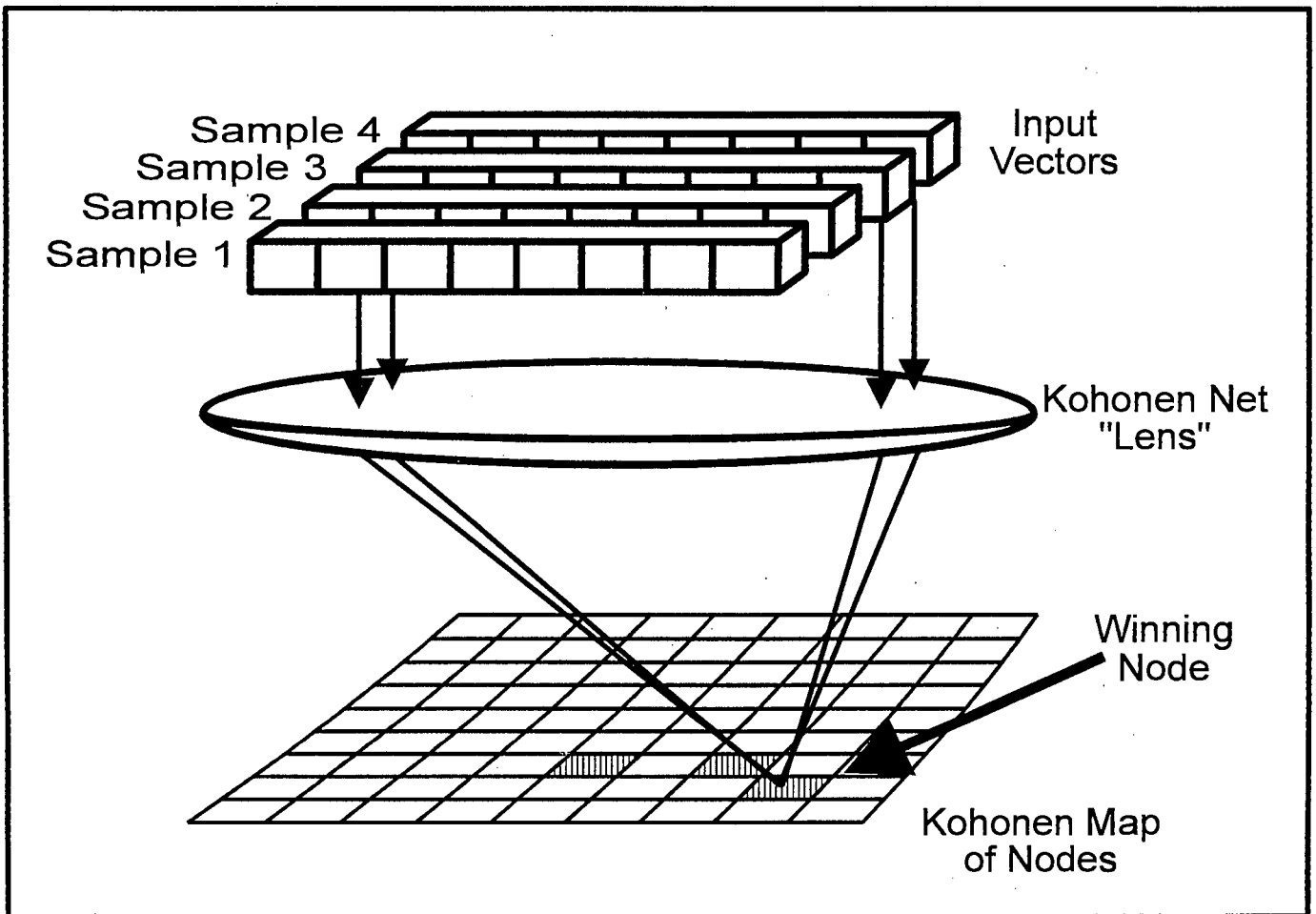


Figure 2.2 : The SOM using the optical lens analogy. Input vectors are projected onto the two-dimension node map (Kohonen map).

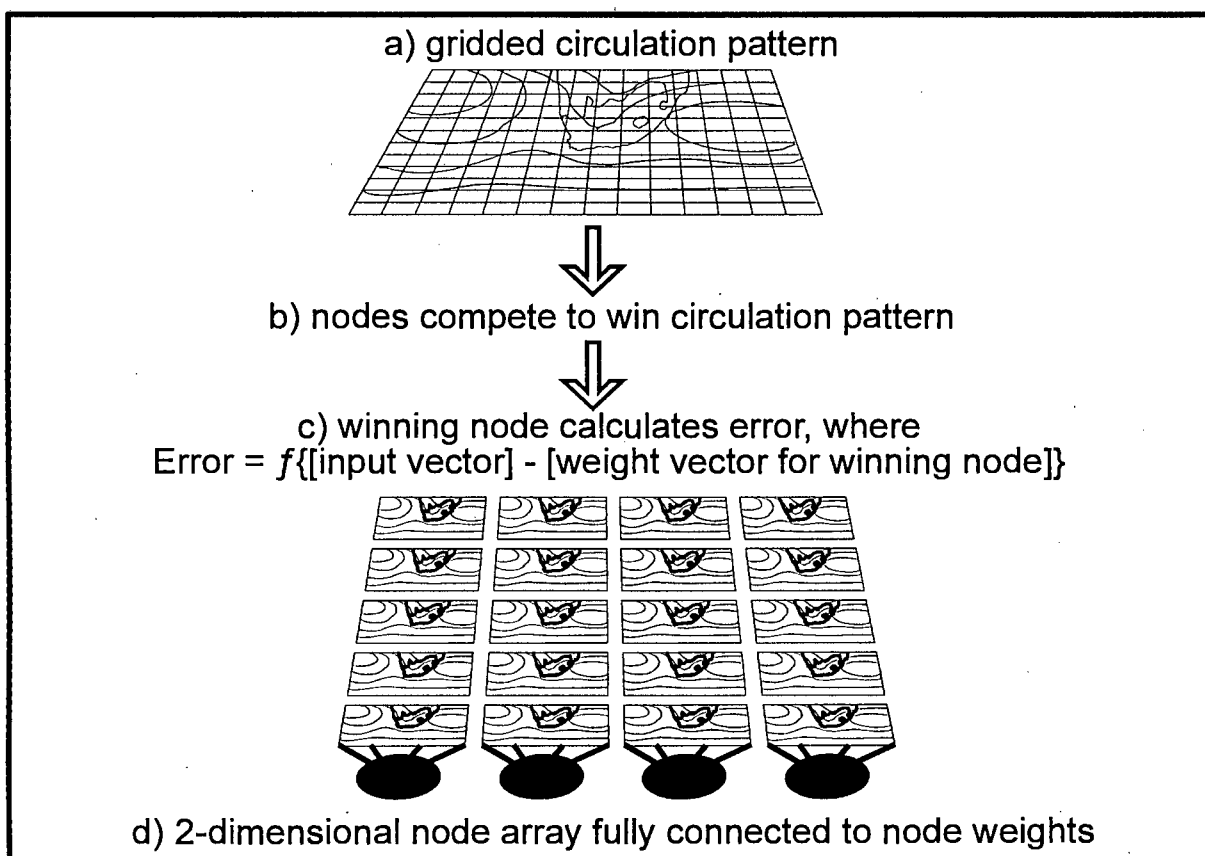


Figure 2.3 : Mapping a circulation pattern to node weights in a trained SOM. An input gridded circulation pattern is presented to the SOM (a), the nodes compete to win the pattern by having the best match (b), the winning node calculates an error (c) and the location of the associated node on the 2-dimensional map is recorded (d).

The combination of input to node mappings, node co-ordinates and associated error measures form the basis of further analysis, and are termed the visualisation data. These data refer to frequency of input patterns on a node, average error and trajectories in node space of the input data sequence. The weights, on the other hand, represent a single circulation map at each node, the archetype for a point in the continuum, and may be plotted as an indicator of average pattern associated with each node. These represent the characteristics of the training data set with respect to circulation patterns. In essence, these weight maps are those which typify the synoptic category. The combination of this information is used for the interpretation of synoptic events and their variability.

The SOM techniques are particularly useful for obtaining a generalised finger-print of the characteristics of circulation patterns in terms of inter-annual and intra-annual variability including the frequency of particular synoptic types. The advantage of this in climate change research is two-fold. Firstly the procedures offer a new method of validating GCM simulations against observed data, and secondly, it allows an investigator to compare experiment and control GCM simulations, and analyse the differences in sequence and frequency of synoptic events.

2.2 GCM Simulations

The General Circulation Model (GCM) is a derivative of weather forecast models which were first initiated *circa* 1956. The purpose of the GCM is to simulate general *climate* behaviour as opposed to weather forecasting. The principal difference being that a weather model requires initial boundary conditions upon which it is sensitively dependent and thus

limits forecasts to approximately two weeks, whereas a GCM aims to simulate the general behaviour of the weather, i.e. the climate. Modelling requires an understanding of the processes taking place within the climate system, and GCMs represent the span of the limited knowledge of the global system. The models are based on the physical laws of fluid dynamics and thermodynamics and solve the equations governing the conservation of energy, mass and momentum. Current generation GCMs use approximately 18 layers in the vertical atmosphere with a horizontal resolution of 3°-4° of longitude and latitude.

While 3-dimensional GCMs began with a purely atmospheric focus, present generation models incorporate sophisticated representations of most components of the hydrosphere, cryosphere and biosphere. Currently, however, the weakest links in the system are, firstly, that of oceans models (Houghton *et al.*, 1992), which is a critical component of climate forcing and, secondly, the difficulties in modelling processes that occur at the sub-grid-scale. With regards to the latter point, the spatial resolution of a model is one of the most important issues in modelling. Processes that operate on a scale smaller than the grid cell cannot be directly represented, and require that they be parameterised in some other form. These include important regional scale processes and elements such as precipitation, convection and cloud processes, vegetation, soil, snow and sea-ice. Parameterisation of these processes ranges in complexity. For example, clouds can be represented by as little as two formations, convective or large-scale stratus clouds which are either present or absent in any of the atmospheric levels in any of the grid boxes. At the coarse scale of a GCM, such generalisations result in the important radiation and precipitation processes becoming grossly simplistic. As a result, precipitation is described as a grid average and seldom represents the temporal or spatial complexity of surface patterns associated with

regional climates, let alone the patterns associated with complex surface topography.

Similarly, soil, vegetation and snow have been typically modelled in a coupled land surface model which may contain, for example, a two layer vegetation model and simple 'bucket' hydrology to resolve surface processes such as runoff, evaporation and percolation. Sea ice is also often modelled separately to try to resolve the complex interaction between ocean, sea-ice and air with respect to water salinity, temperature and albedo. Oceans are particularly important components in driving the atmosphere and yet may be represented in GCMs as a simple 'swamp model', which acts only as a source of water vapour, or as a single mixed layer, and only seldom as a full ocean model. Due to the difficulties associated with a full coupled ocean-atmosphere model, namely differing time scales of operation and computational requirements, the 'swamp' and mixed layer ocean parameterisations are more common (Peixoto and Oort, 1992). However full ocean models coupled to the atmospheric GCMs are being developed and hold significant promise for improving GCM simulations (Murphy and Mitchell, 1995).

Despite some of the caveats, the GCM is capable of simulating many of the large scale features of the observed climate. However, the value represented by a GCM grid block is typically an areal average of the grid cell. As a result, values tend to be smoothed representations of a process and thus may affect interpretation of results.

In order to investigate the impact which an increase in CO_2 has on the climate, a control and experimental simulation are typically performed with the GCM, where the boundary and initial conditions are identical in both simulations, except that in the experimental simulation the CO_2 concentration is doubled. By comparing the results of the two

simulations the impact of the increase in CO₂ can be determined. A climate change experiment can be conducted by either increasing atmospheric CO₂ levels in the model at about 1% per year the *transient* approach, or by re-initialising the model with the same boundary conditions and instantaneously doubling the atmospheric CO₂ - the *equilibrium* approach (for example Mitchell *et al.*, 1995). Climate change scenario development is principally based on data from such GCM simulations, with consequent difficulties at the regional and grid scales.

Validation of GCM data is important in giving some credibility to a GCM derived climate scenario. This typically involves comparing the general behaviour of simulated circulation from the control run with present day observed circulation patterns. This validation procedure attempts to identify disparities or incongruencies in the model, and provides a basis for the confidence level which may be placed on the climate change scenarios.

Approaches to validation differ, but, on the whole, the processes are generally focused on a simple statistical description of the means and variances of the climatic variable of interest, where values are checked for consistency with the observed data (Santer and Wigley, 1990; Willmot, 1982; Willmot, 1981).

A significant proportion of GCM work revolves around validation of GCMs. For example, the Atmospheric Model Intercomparison Programme (AMIP) focuses on comparing the physics and parameterisation characteristics of various GCMs, with the aim of providing a comprehensive assessment of the capability of models to represent mean seasonal states and large-scale inter-annual variability. The project involves 30 modelling groups conducting model simulations for the period of 1979-1988 using observed SST and sea-ice

distributions and standard values for the various boundary conditions (Gates, 1992).

Validation of the general modes of behaviour is useful procedure when applied to GCM output and one to which synoptic climatology analysis lends itself.

Once validated, and assuming the control run is realistic, GCM output may then be utilised, in the context of developing climate change scenarios, by analysing the difference between the present day simulated control run and the doubled CO₂ simulated environments.

However, while valuable information can be derived at continental and hemispheric scales, regional information is still subject to the grid cell data limitations of GCMs. Evidently, GCMs are better suited to simulating the general synoptic scale circulation, and due to the restrictions in sub-grid-scale modelling, validation of the circulation is critical.

2.3 Summary of Procedures

In summary, the Kohonen SOM is a potentially valuable technique for GCM studies, and one that perhaps offers more insight than more traditional synoptic climatology techniques. Furthermore, the SOM categories may be depicted in terms of typical circulation patterns and each observation may be expressed in terms of a location on the continuum of SOM categories. Frequencies and trajectories within the two-dimensional space associated with intra-annual and inter-annual variations highlight seasonal characteristics of the circulation and are useful for characterising a particular climate. The aims and techniques of the SOM closely match those of synoptic climatology, which is concerned with relating large scale circulation to the surface environment through classification of circulation patterns. The study of the characteristics of a climate is highly pertinent to climate change studies, in which anthropogenic influences are likely to impact on the climate, although the exact

manifestation of climate change is not fully known.. General circulation models are probably the most effective tools to use for investigating future climates, however there are a number of caveats as a result of the constraint of computing capacity, particularly the lack of sub-grid-scale information. Nevertheless, GCMs are currently able to simulate large scale processes with a high degree of skill and hence their attractiveness for use in synoptic climatology based climate change studies.

CHAPTER THREE

Data Preparation and Methods

3.0 Study Area and Data Preparation

In order to study circulation patterns over southern Africa a suitable synoptic window needs to be chosen that encompasses the atmospheric circulation which has direct effects upon the climate of the region. While regional climates are influenced by spatially remote features in the climate system, these influences are, by the nature of the interconnectedness of the climate system, represented in the circulation patterns over the region. Given that synoptic scale processes take place on a scale in the order of 30° of longitude or latitude, a synoptic window was defined with boundaries extending from $15^\circ 00'W$ to $65^\circ 00'E$ and $10^\circ 00'S$ to $50^\circ 00'S$. In order to represent the atmospheric circulation, sea level pressure (SLP) values are used.

Three data sets were made use of for this application of the SOM, they are the Australian Southern Hemisphere observed sea level pressure data set and the control and the doubled carbon dioxide simulations from the GENESIS version 1.02 GCM.

Observed sea level pressure (SLP) data were extracted from the Australian Southern Hemisphere data set. The data are the result of a gridded analysis of sea level pressure produced by the Australian Bureau of Meteorology and the methods used to obtain the data are documented by Le Marshall *et al.* (1985). These data are derived using observed values which are gridded onto nominally equidistant points (47×47) and overlaid on a polar

stereographic projection (Jenne, 1975). The data set is twice daily (0Z and 12Z) and spans 17 years from 1973 to 1989.

GCM data were extracted from the control and doubled carbon dioxide ($2\times\text{CO}_2$) simulations of the GENESIS v1.02 GCM. The GENESIS GCM comprises an Atmospheric GCM (AGCM) with a spectral resolution R15 grid (approximately 4.5 degrees of latitude and 7.0 degrees of longitude), 12 levels of atmosphere with three possible cloud types, a $2^\circ \times 2^\circ$ land surface model, multi-layer models of vegetation, soil, snow, sea ice and a 50 metre slab ocean layer (Pollard and Thompson, 1995). The data are twice daily SLP (0Z and 12Z) and span 5 GCM years.

Since the purpose of the analysis is to compare the data sets, namely, the observed with the control run and the control run with the doubled CO_2 , the data sets are then prepared onto a common spatial grid and temporal scale. The R15 grid of the GCM data is the lowest resolution grid, thus the observed SLP data were re-interpolated to the lower resolution R15 grid (Figure 3.0). As a precaution, extreme outlying values of SLP for the Southern Hemisphere data set are removed before interpolation to the R15 grid. This process requires calculating the standard deviation for all SLP values through time and space and removing the entire set of observations for a particular day in which one of the values lies outside the sixth standard deviation boundary with respect to the mean of the entire data set. Removal of these extreme observations minimises skewing of SOM results. This resulted in 0.46% being removed from the observed data, 0.11% from the GCM control simulation and 0.37% from the GCM doubled CO_2 simulation.

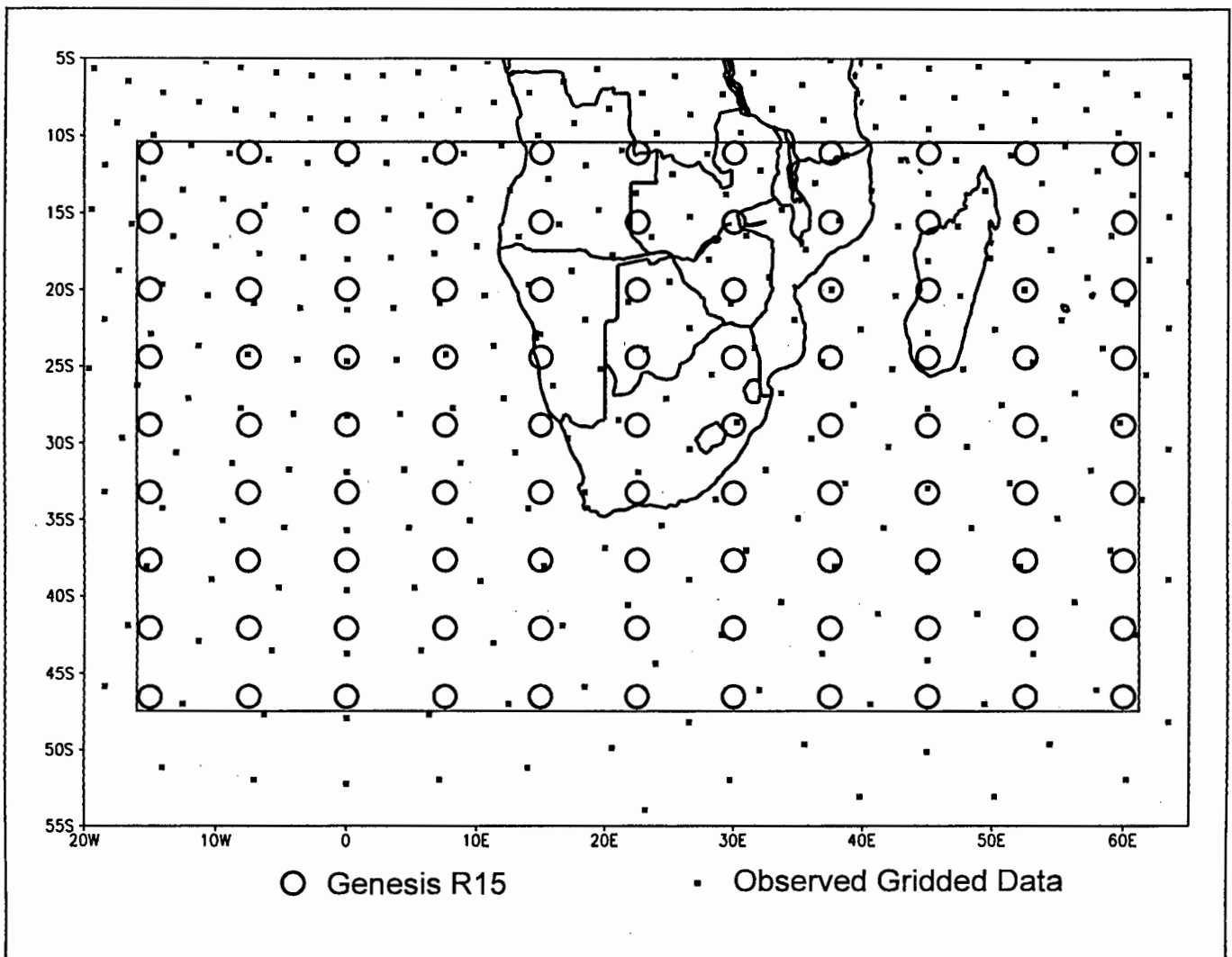


Figure 3.0 : Map of study area (inner box) and layout of the GENESIS R15 grid points (circles) and the Australian Southern Hemisphere observed gridded data set (squares).

The re-gridding scheme used is a spherical interpolation algorithm suggested by Wilmott *et al.* (1985). The spherical interpolation algorithm uses inverse square distance interpolation with spherical distances, and accounts for directional bias in the source data. Interpolation introduces a maximum error of approximately one millibar. The algorithm uses a defined search radius, and a minimum and maximum number of points to interpolate from. In this application the search radius was set to 7° , and a minimum of three points, and a maximum of nine points were used in interpolating each point. The final target domain is defined by 15°W ; 10°S to 65°E ; 50°S (Figure 3.0) and comprises 11 grid points of longitude and nine grid points of latitude, a total of 99 grid points.

Monthly mean data sets were then calculated from the daily data. This decision was guided by characteristics of commonly used GCM output data, in particular those of the Atmospheric Model Intercomparison Programme (AMIP). AMIP consists of an agreed standard with which different GCMs may be compared, and an aspect of the programme derives monthly statistics for various variables from models which are run in 'climate mode'. Thus, in the context of available AMIP data, the equivalent observed and GENESIS GCM data were created, and using the respective daily data sets, monthly means of observed SLP, control run SLP and doubled CO_2 SLP are calculated, each with 6026, 1882 and 1877 observations respectively. This generated 201 observations of monthly means for the gridded observed data, and 62 observations for each of the control and $2\times\text{CO}_2$ GCM simulation data sets.

The final step in the data preparation is to standardise the three data sets. Standardisation is useful for comparing data sets where the means have a bias. While this forces the means of the GCM data to that of the observed data, this was considered acceptable in the light of

the objective of focusing on spatial pattern. If the mean were retained this would obscure the process of synoptic pattern comparison between observed and GCM data.

Two approaches to standardisation were considered, each varying the final output and hence the interpretation of the results; matrix-wise or column-wise standardisation, where the data matrix is configured with each column forming a time series for a particular grid point. Matrix-wise standardisation standardises each data point in terms of the mean of the entire grid through time, while column-wise standardisation normalises each grid point time series in terms of the mean for that column. Matrix-wise standardisation thus expresses each grid point in terms of standardised departures from the long-term areal mean, while column-wise standardisation expresses each grid point in terms of standardised departures from the long term mean of that grid point. For the purposes of this study a matrix-wise standardisation is used so as not to distort the spatial continuity of the data.

Standardisation also has the effect of removing overall differences in magnitude of pressure between two data sets when comparing them. Table 3.0 illustrates the overall difference of approximately four millibars in the mean of the observed SLP and the two GCM SLP data sets, while the variabilities of the data are comparable. This is important, for example, when comparing a control run and experiment run where there might be net increases or decreases in pressure values in the area under consideration. The effect of standardisation thus needs to be kept in mind during the analysis, remembering that it does not affect the seasonality of the circulation patterns, just their overall magnitude.

Data set	Dimensions	Mean (hPa)	Std Dev. (hPa)	Duration (monthly)
Observed SLP	201 x 99	1014.7	5.054	1973-04-15 to 1989-12-15
Genesis control run	62 x 99	1010.8	5.411	0001-01-15 to 0006-02-15
Genesis doubled CO ₂	62 x 99	1010.6	4.952	0001-01-15 to 0006-02-15

Table 3.0 : Summary statistics of the three circulation (SLP) data sets before standardisation. Mean and standard deviation are calculated with respect to the entire data set (matrix-wise).

3.1 Methods and Procedures of the SOM

This section outlines possible variations in the procedure for using the Self-Organising Map (SOM). Alternatives presented here are based on those available within the SOM_PAK software system.

3.1.0 Structural Considerations

The two-dimensional array of nodes onto which the images of each input vector will be projected comprises both size and shape. The shape, or topology, may be hexagonal or rectangular, and refers to the way in which nodes in the node space are arranged, as shown in the Figure 3.1. The hexagonal topology provides a more effective intra-node proximity and hence provides a slightly more continuous image than the rectangular topology (Kohonen *et al.*, 1996). The SOM_PAK manual suggests no significant advantage of the hexagonal topology, apart from visualisation advantages, hence the rectangular structure (Figure 3.1 b) was preferred, for ease of use in the analysis.

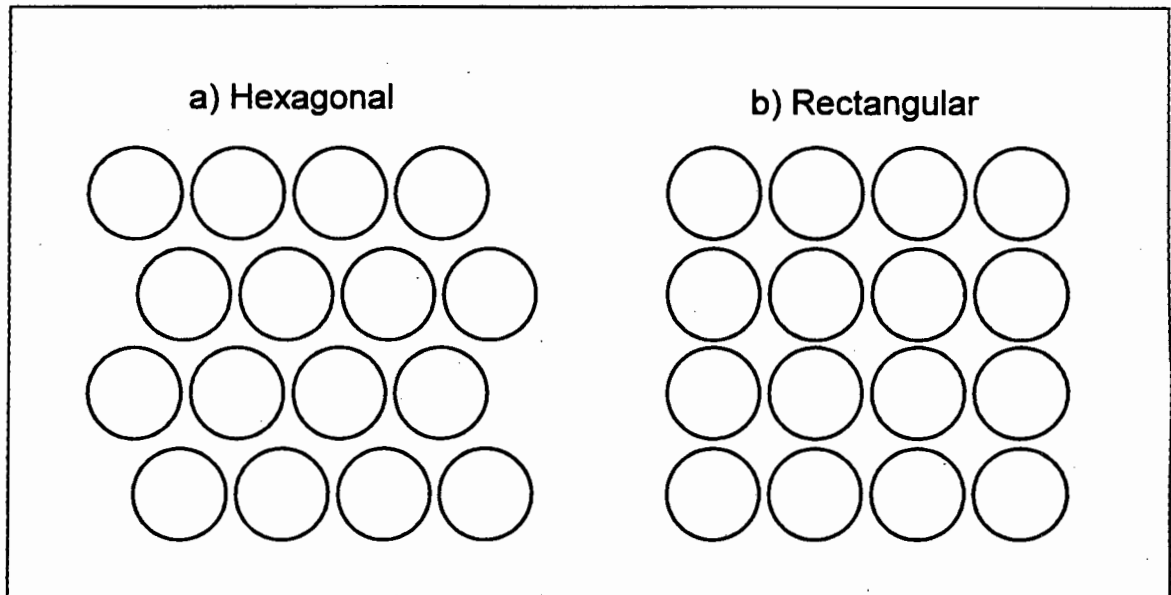


Figure 3.1 : Two possible topologies a) hexagonal and b) rectangular, that may be selected for the two-dimensional node map or SOM.

Determining the size dimensions of the target domain requires a slightly heuristic approach. The SOM weights, once the SOM is trained, represent the probability density function of the input data, where the probability density function is the frequency of the observations in the high ($11 \times 9 = 99$ grid points) dimensional input space. Thus, the elastic network formed by the weights must be orientated along with the probability density function (Kohonen *et al.*, 1996). The only guideline is that the shape of the SOM has some stable orientation, that is, the shape must not be perfectly symmetrical. Thus, a rectangular domain as opposed to a square or circle, should be used (Kohonen *et al.*, 1996). Aside from this constraint the researcher may determine any reasonable dimensions for the node space which ultimately determine the maximum number of possible categories into which the input data set may be divided. In this vein, the form of the array dimensions should correspond to the major dimensions of the probability density function. Thus, the greater the size of the dimensions, the greater the number of archetypes that may be resolved across the continuum of samples, at the expense of less generalisation of synoptic types.

Conversely, if the size of the dimensions is small, the data is largely generalised with only the primary and dominant synoptic states identified. The extremes of the dimension would thus produce, with a single node, some representation of the mean, or with as many nodes as samples, a one-to-one mapping between input and output. Ultimately, however, the decision is a subjective one, and in this study the oblong dimensions of four by six nodes was selected, allowing for a total of 24 synoptic types.

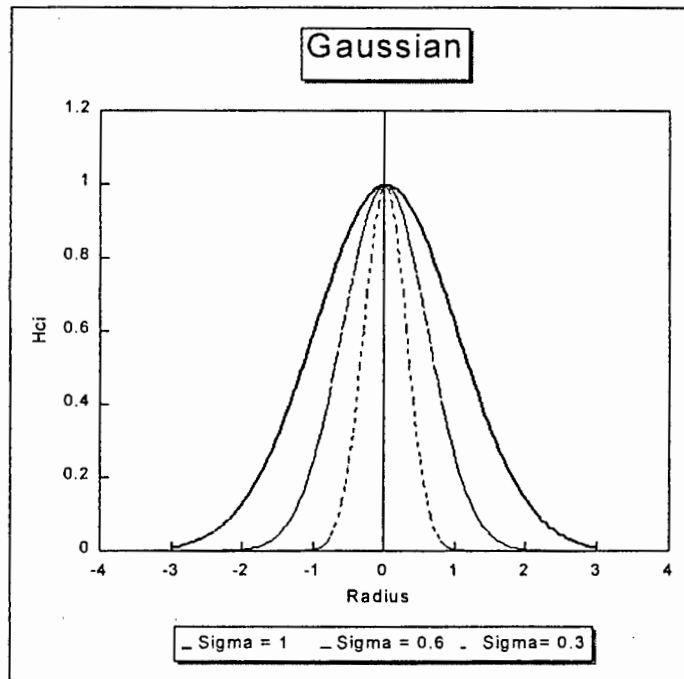
3.1.1 Training Considerations

Training is implemented with one of two possible neighbourhood *kernels* which are central to the SOM algorithm, *bubble* or *Gaussian*, and relate time and learning rate together to determine how a particular node's weights are updated such that,

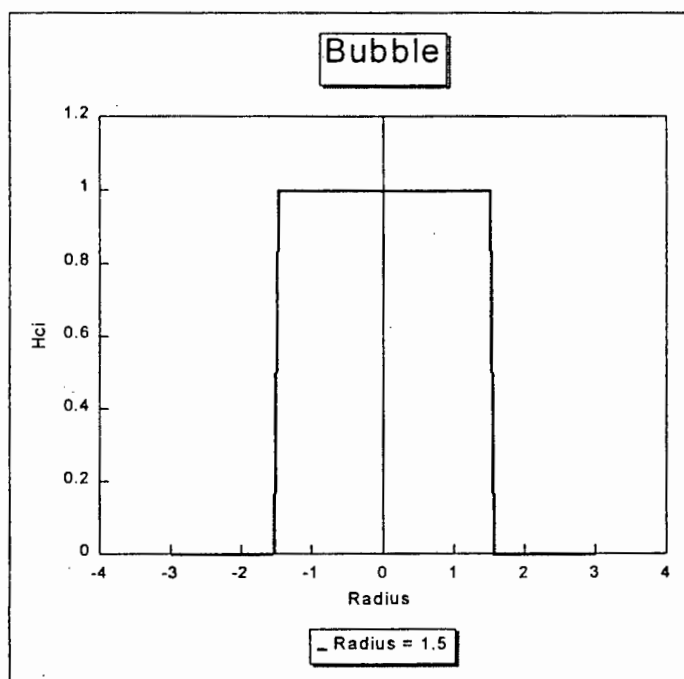
$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)], \quad (1)$$

where m are the weights, x is an input vector, t is the time step integer and h_{ci} represents the neighbourhood kernel (Kohonen, 1990).

The bubble neighbourhood kernel evaluates the weights only if they are in the update neighbourhood, defined by a radius, with respect to the winning node. Weights outside that area are left intact. Similarly, the Gaussian neighbourhood kernel updates weights within a radius of the winning node except that the update rate increases toward the centre of the radius, approximating a 'bell curve'. A second variable, sigma (σ) determines the shape of the bell curve and is reduced as a function of time during training (Figure 3.2). Both kernels provide a mechanism for the radius of the update neighbourhood to decrease



a)



b)

Figure 3.2 : a) The Gaussian neighbourhood kernel updates weights within a radius of the winning node except for the update rate increases toward the centre of the radius, approximating a 'bell curve', this figure shows three variations of the structure determined by sigma. b) The bubble neighbourhood kernel evaluates the weights only if they are in the update neighbourhood, weights outside that area are left intact.

with time. In addition, the kernels are functions of a learning rate (α), such that $0 < \alpha(t) < 1$. The learning rate determines how easily the weights may be changed. In general, the learning rate becomes smaller towards the end of training so that spurious weight configurations do not affect the overall generalisation represented by the weights. Two schemes are available for reducing α over the course of the training, being *linear* and *inverse_t*. The linear α reducing scheme simply reduces α linearly to zero over the course of the training while the *inverse_t* method relates α and time in an inverse relationship (Figure 3.3).

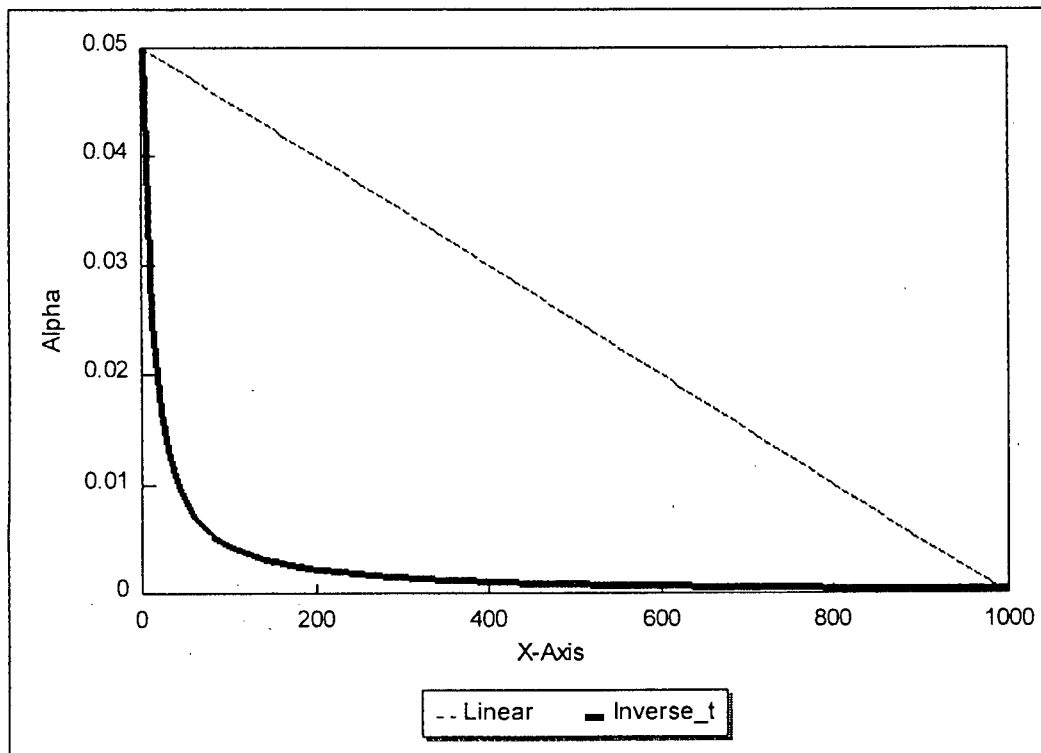


Figure 3.3 : Graph showing the way in which the learning rate (α) is reduced with the linear and *inverse_t* schemes over 1000 iterations.

Determining the number of iterations for training has been suggested to equal, as a rule of thumb, 500 times the dimensions of the node array. Training may vary from up to 100,000 steps down to around 10,000 for fast learning (Kohonen 1990). In the light of the

dimensions used in this study it was decided subjectively that the rule of thumb value of 12,000 ($4 \times 6 \times 500$) was too near to the fast training limit and thus a value of 30,000 was chosen.

The SOM_PAK package offers two means by which the efficacy of training may be judged, the first being the *Sammon mapping* and the second, the set of holistic error calculations or quantisation errors. A Sammon mapping comprises a map of the node inter-relations in weight space with respect to their Euclidean distances (Sammon, 1969), and is the two-dimensional projection of the probability density function of the n -dimensional input data. These mappings are useful for determining the relationship and continuity across the two-dimensional space and may throw some light onto specific interpretations of the data. The Sammon mapping may be visualised by plotting its projection onto a two-dimensional plane. The plot serves as a generalised view of the inter-relationship of each node with respect to each other.

Two measures of error are available, the *average quantisation error* and the *average distortion measure*. These calculate an overall measure of error associated with an input data set and a set of weights. An average quantisation error calculation attempts to express, in a single real number, the accuracy of SOM by calculating the average of all errors between each input vector and that vector's best matching node, while the average distortion measure uses a weighted distance measure dependent upon the neighbourhood function, viz. bubble or Gaussian. The average distortion measure is however especially useful with larger maps, and thus the simpler average quantisation error was used for the purposes of this study. Training is deemed complete on achieving some threshold minimum of error or when the incremental decrease in error becomes small. If the weights

are initialised randomly, then to ensure an adequate training, a number of trainings should be run to determine the optimum subjective balance between a good Sammon mapping and a low quantisation error. Provision is made in the package to take *snapshots* copies of the weights at set intervals during training, the quantisation error may then be calculated for each snapshot and a graph of progress may then be plotted.

3.1.2 Training

The object of training is to minimise the overall error. There are two main choices that may be made to determine how training is executed. The first relies on a random initialisation of the weight space and the second relies on an ordered weight space that is initialised with the first two orthogonal components (as calculated with principal components analysis) of the input data set opposed about the geometric diagonal of the node space.

3.1.2.0 *Random Initialisation*

In this training scheme the weight space is simply initialised by a set of random numbers. Training then proceeds in two phases, first a coarse sorting phase in which the learning rate is high and the update neighbourhood is large relative to the dimensions of the node space. This phase of the training has the effect of general ordering of the node space with respect to the weight characteristics of each node. Secondly, the 'annealing' phase, which can be likened to the annealing process in crystallisation, is run for a large number of iterations, a low learning rate and a small radius of influence on the weight space. This second phase allows more local differentiation between nodes.

3.1.2.1 *Linear Initialisation*

In this scheme the weight space is initialised with the first two orthogonal components of the input data set which are calculated by a principal components analysis (PCA). The loadings of the first two components are used to initialise the weights with each component opposed about the geometric diagonal of the node space. This is achieved by setting the values of the weights in one corner of the node map to the loadings of the first component produced by PCA reduction of the input data set. The same is repeated for the opposite corner using the second component of PCA reduction of the input data set. The remaining intermediate nodes of the node map are initialised to various combinations of the two components related to their relative distance from each corner, thus creating a continuum in node space between the two PCA components. Training that follows is an annealing process in which the learning rate is low and the radius of influence is small. This is the preferred method in this study, as the primary mode of variance in the circulation data sets tends to be the seasonal cycle, this allows the opposing phases of the annual cycle to orientate at each extreme of the stable node space. The random initialisation does not ensure an orientation on the basis of the seasonality. Verification of this may be seen by decomposing the weights in the analysis stage.

3.1.3 **Post-Training Analysis and Visualisation**

Once training is complete two new data sets are created, one containing the weights associated with each of the target nodes, and the second containing observation-to-node mappings and their associated error.

3.1.3.0 *Decomposing Weights*

Each set of weights that represent a node may be used to illustrate the structure of the weights. This is made possible by the fact that at a node the weight vector may also be expressed as a corresponding weighting associated with each input grid point. The result is that the weights may be plotted on a latitude and longitude map. These maps represent the spatial pattern of standardised sea level pressure that characterise each node and thus each SOM category. When all such maps are plotted and arranged in the same order as the nodes that they represent, a node map of weight maps or *meta-map* is constructed. The meta-map is the key to interpreting subsequent patterns derived from the SOM analysis.

3.1.3.1 *Decomposing Nodes*

After training, each observation from the input data set is presented to the SOM, which then determines a winning node on the basis of the smallest error from the error function between the weights of the winning node and the input values of the observation (Figure 2.3). The node co-ordinates, to which each sample vector is mapped on the two-dimensional node map, are recorded along with error. These data represents the list of input observations and their corresponding node, with associated error, as determined by the SOM. The data forms the basis of further analysis of variability on the seasonal and annual time scales.

Using the above diagnostic data, further analyses of the temporal nature of the data may then be conducted. These include plotting the trajectory of the time series in the node space, compositing the node mapping month (for intra-annual variability), plotting the frequency dispersion in node space and plotting the average error in node space. Diagrams representing these analyses are presented in the results.

CHAPTER FOUR

Application of the SOM

4.0 Application of the SOM

This section will illustrate the evaluation of circulation modes using monthly observed circulation data from the Australian Southern Hemisphere data set and GENESIS GCM simulated control and doubled atmospheric CO₂ circulation data. The procedure entails deriving the characteristics of observed monthly circulation using the classification capabilities of the Self-Organising Map (SOM). If the GCM control simulation data show acceptable similarity to the observed data, validation, then the 1xCO₂ data are compared with GENESIS GCM 2xCO₂ simulation. Key differences, as highlighted by the SOM, are used to interpret likely changes expected in a double carbon dioxide environment.

Two SOMs were trained, one using the observed monthly data and the second using the GCM 1xCO₂ circulation data. In each case a SOM comprising 24 nodes (4 in the x-axis and 6 in the y-axis) was initialised using the *lininit* option, which orders the weights with respect to the two principal eigenvectors of the input data, as described in the previous chapter. For visualisation purposes a *rectangular* topology was used. The training *kernel* was configured to use the *bubble* neighbourhood and an *inverse_t* method of adjusting the learning rate. Once initialised the training was run for 30,000 iterations, with an initial *learning rate* (α) of 0.1, during which samples of the weights (*snapshots*) were stored every 500 iterations to monitor the training. A total of 20 complete trainings were

conducted from which the most suitable training was selected. The best trainings are judged as those trainings with the lowest resulting error. From this group of trainings visual inspection of the respective Sammon mappings were undertaken to determine which training to use for the study.

4.1 Observed circulation

A SOM, of the structure described above, was initialised and trained 20 times with the monthly observed circulation data for which there are 201 observations. The 20 Sammon maps were derived from the trained weights and inspected to determine the most suitable training, this being training number two, for which the Sammon mapping shows an even spacing of nodes with respect to their characteristic weights (Figure 4.0). By comparison, Figure 4.1 illustrates a Sammon mapping whose training would have been rejected due to inconsistent spacing and overlapping of weight vectors. Such a configuration would mean that circulation modes that were not adjacent in the input space might be characterised by the same archetype, as represented by the weights. There are a number of solutions for each training and, by way of inspecting the Sammon mapping, the most suitable training is selected. This weight configuration, while valid, would hamper interpretation in later stages of the analysis. A time series of the average quantisation errors during training reveal the way in which the error is reduced through the training (Figure 4.2).

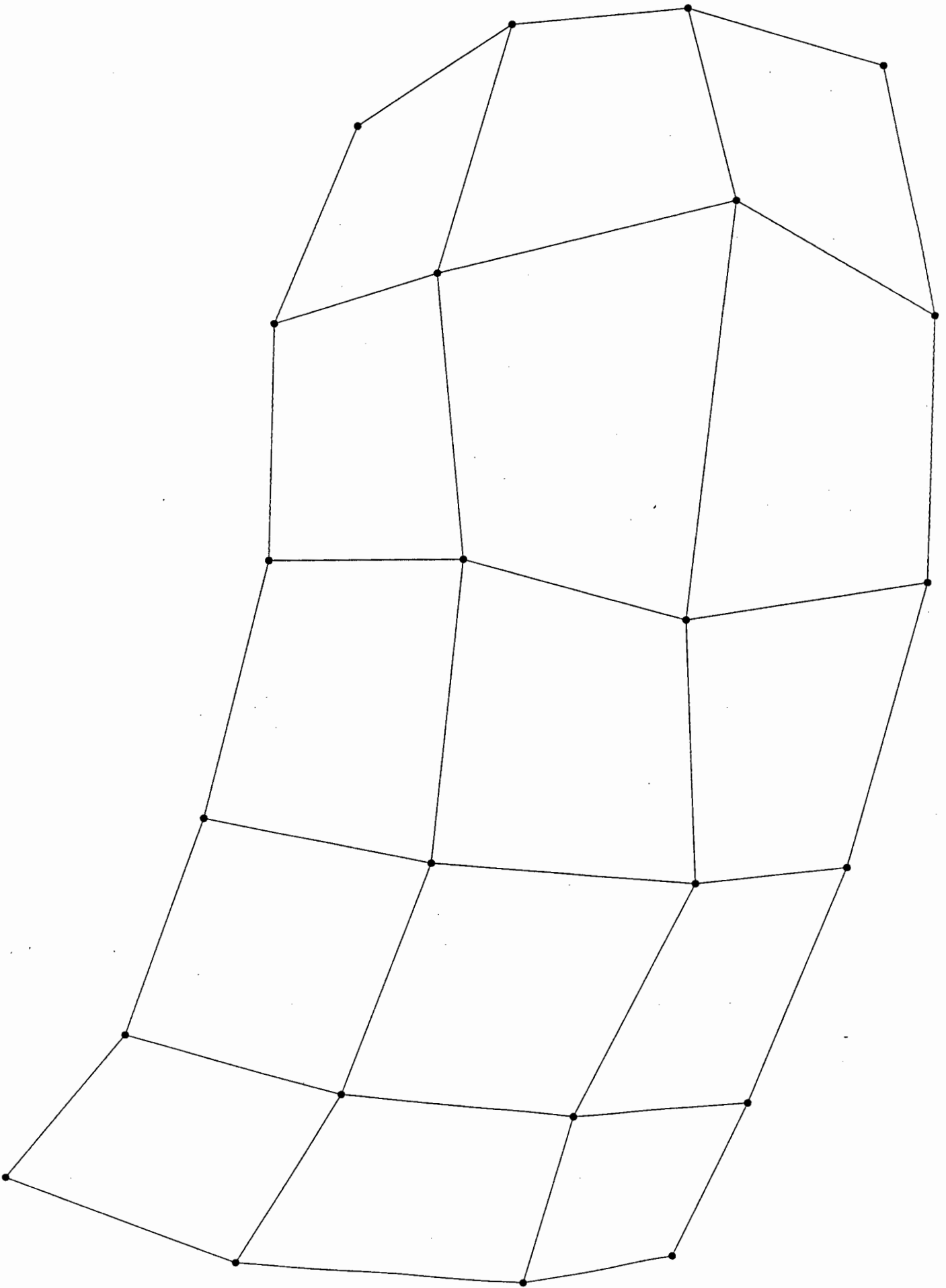


Figure 4.0 : Sammon mapping from the second training with Observed SLP circulation data.

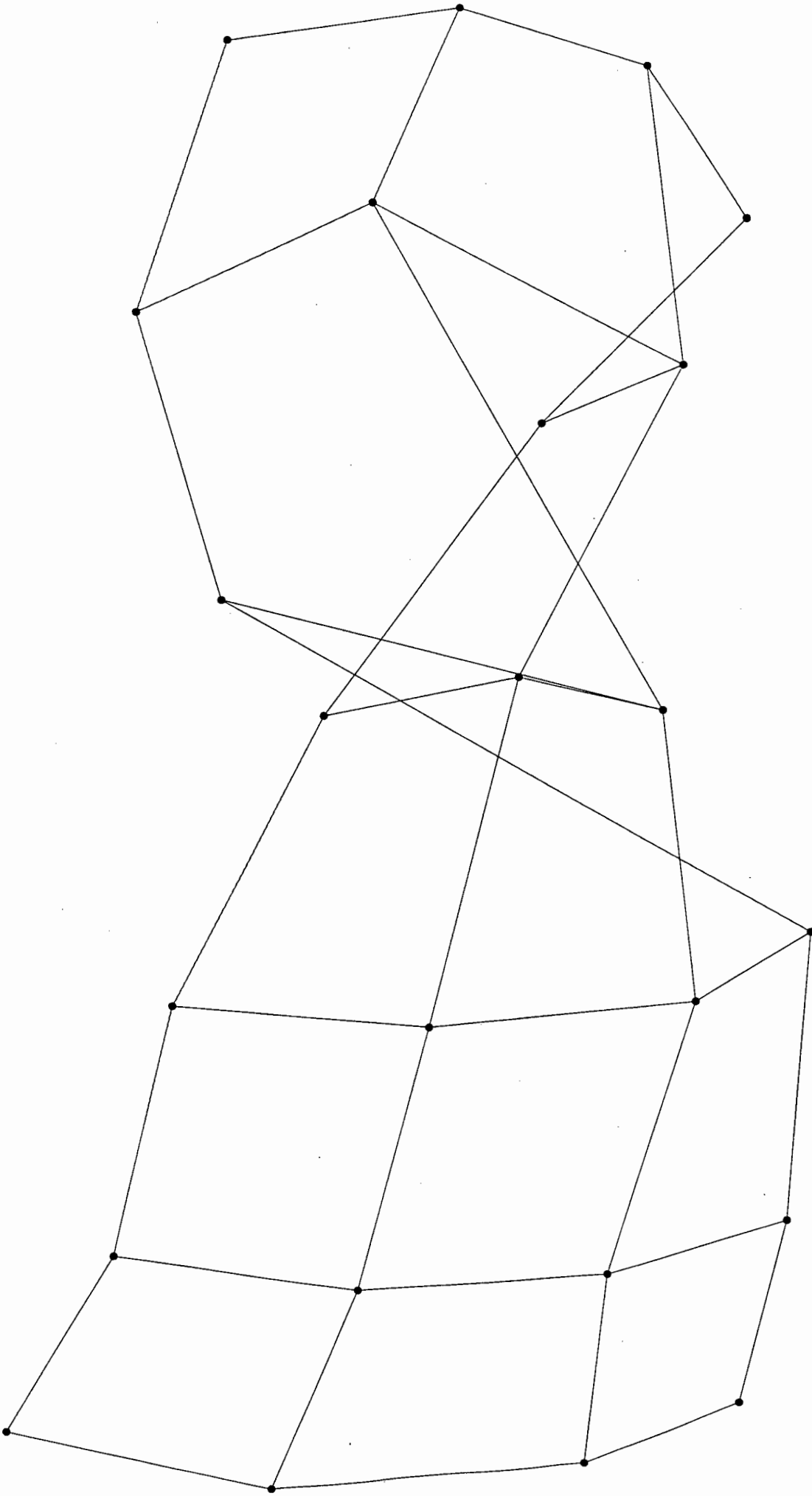


Figure 4.1 : Sammon mapping from the ninth training with Observed SLP circulation data.

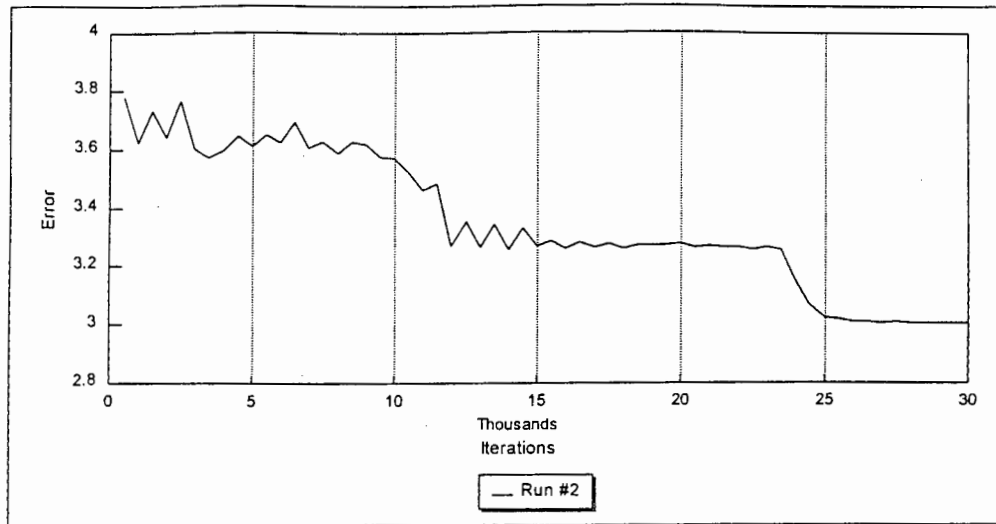


Figure 4.2 : Time series of errors calculated from the snapshots of the weights during training number 2.

Using the trained SOM, a quantisation error was then calculated for the observed SLP and the GENESIS GCM control simulation data ($1xCO_2$) which reveals an average quantisation error of 3.005 and 6.715 per sample respectively. These data indicate the difficulty of matching circulation patterns in the GCM $1xCO_2$ circulation patterns to those in the observed circulation patterns. For example, had the circulation patterns in each data set been similar, one would generally expect less difference between the two average quantisation errors, although this may be skewed by differing frequencies of a particular archetype between each data set.

The weights of the SOM were then analysed by plotting the *meta-map* for the SOM trained with observed circulation patterns (Figure 4.3). The image shows four columns and six rows of maps corresponding to the layout of the two-dimensional node array, i.e. each map corresponds to a node and may be referenced using a co-ordinate system such that the lower left map is (1,1) and the top-right most map is (4,6). Each map represents the typical

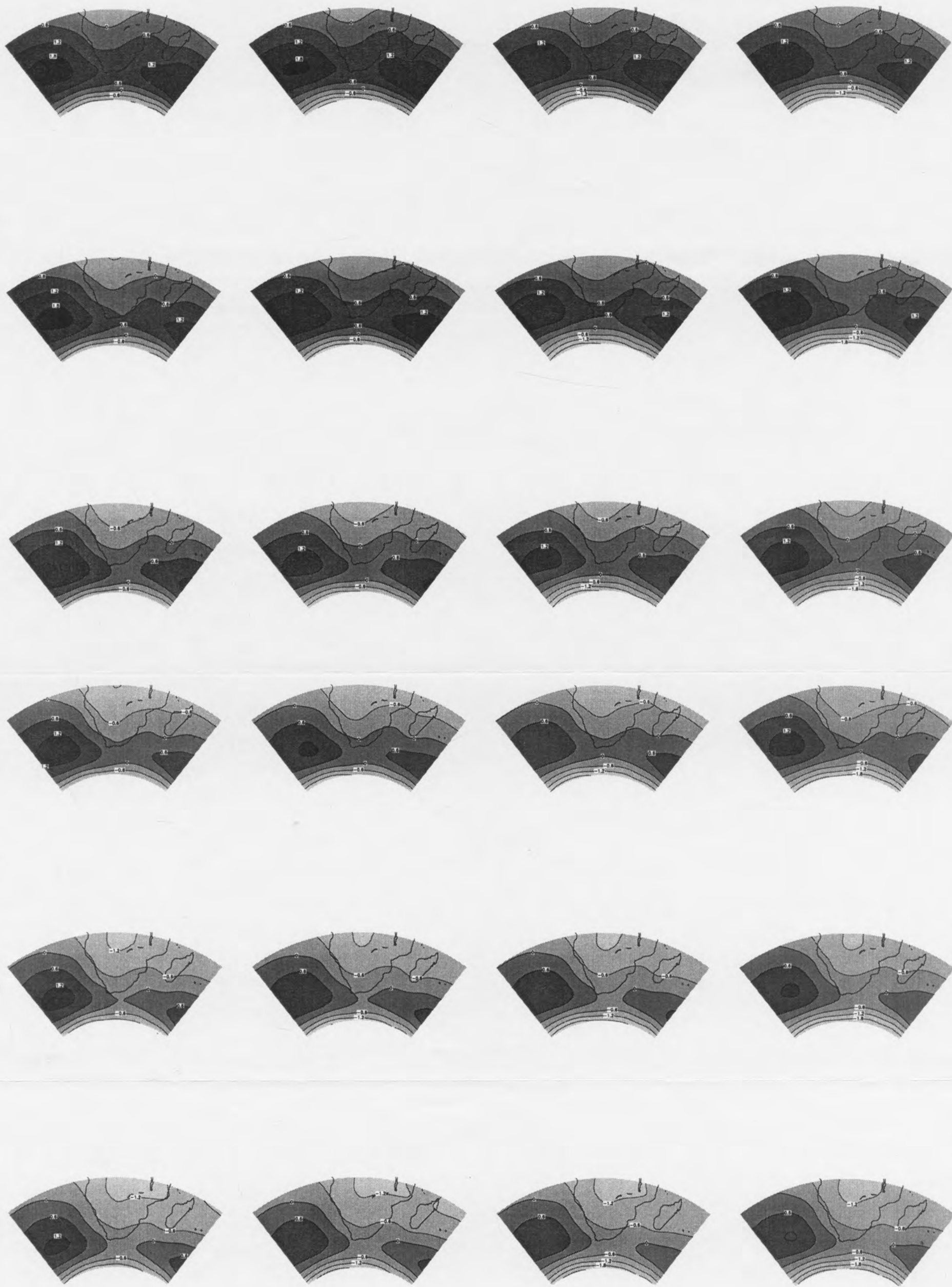


Figure 4.3 : Meta-map for Observed SLP circulation.

map pattern associated with that node and, as can be seen from Figure 4.3, the map patterns are spread in a continuum across the two-dimensional node space. For example, Figure 4.4 illustrates a monthly average circulation pattern for the observation of June 1979 that was mapped by the SOM to the top-right (4,6) corner of the map and which matched that node with the least error of all other mappings to that node. Similarly Figure 4.5 illustrates the observation of January 1974 that was mapped to the bottom-right (1,1) of the SOM map, also with the least error with respect to all other mappings, or hits, to this node.

In general, each map shows varying interactions between the mid-latitude cyclones, the sub-tropical trough and the south Atlantic high and south Indian high pressure systems, with dominance of the sub-tropical trough in the lower part of the meta-map and dominance of the high pressure systems toward the top of the meta-map. The meta-map provides the basis from which to compare and analyse frequencies of synoptic types with respect to inter-annual and intra-annual variability.

The set of observation-to-node mappings were then computed for both the observed and the GCM control run data. These visualisation data are derived by presenting each circulation map pattern to the SOM which maps the observation to the node with the minimum error, and records the error and location of that node.

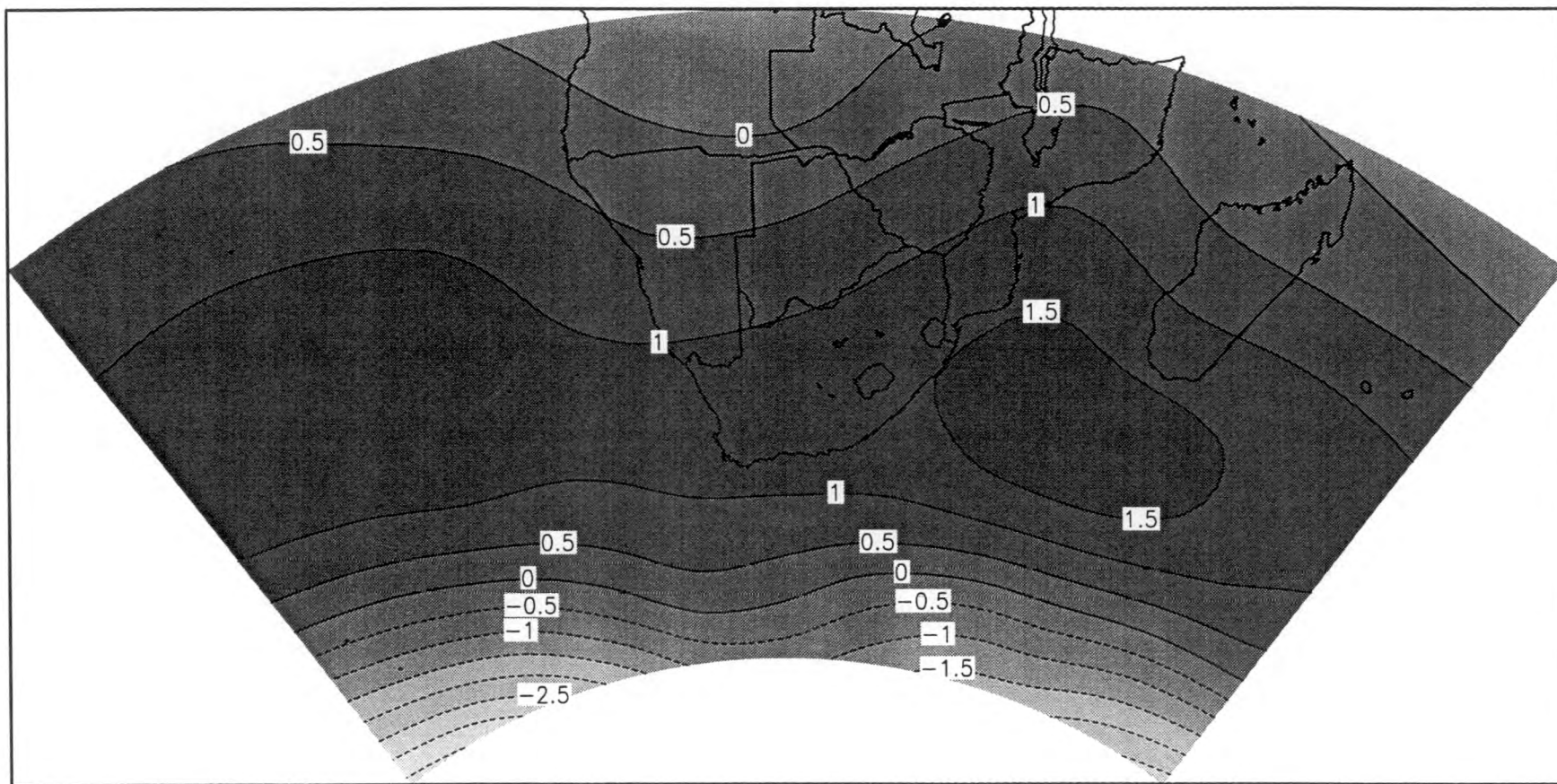


Figure 4.4 : A sample of the best matching input circulation pattern from June 1979 won by node (4,6)

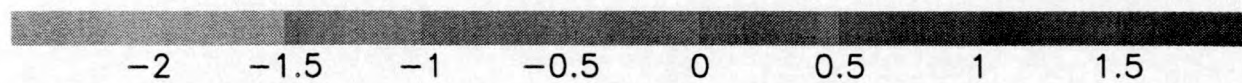
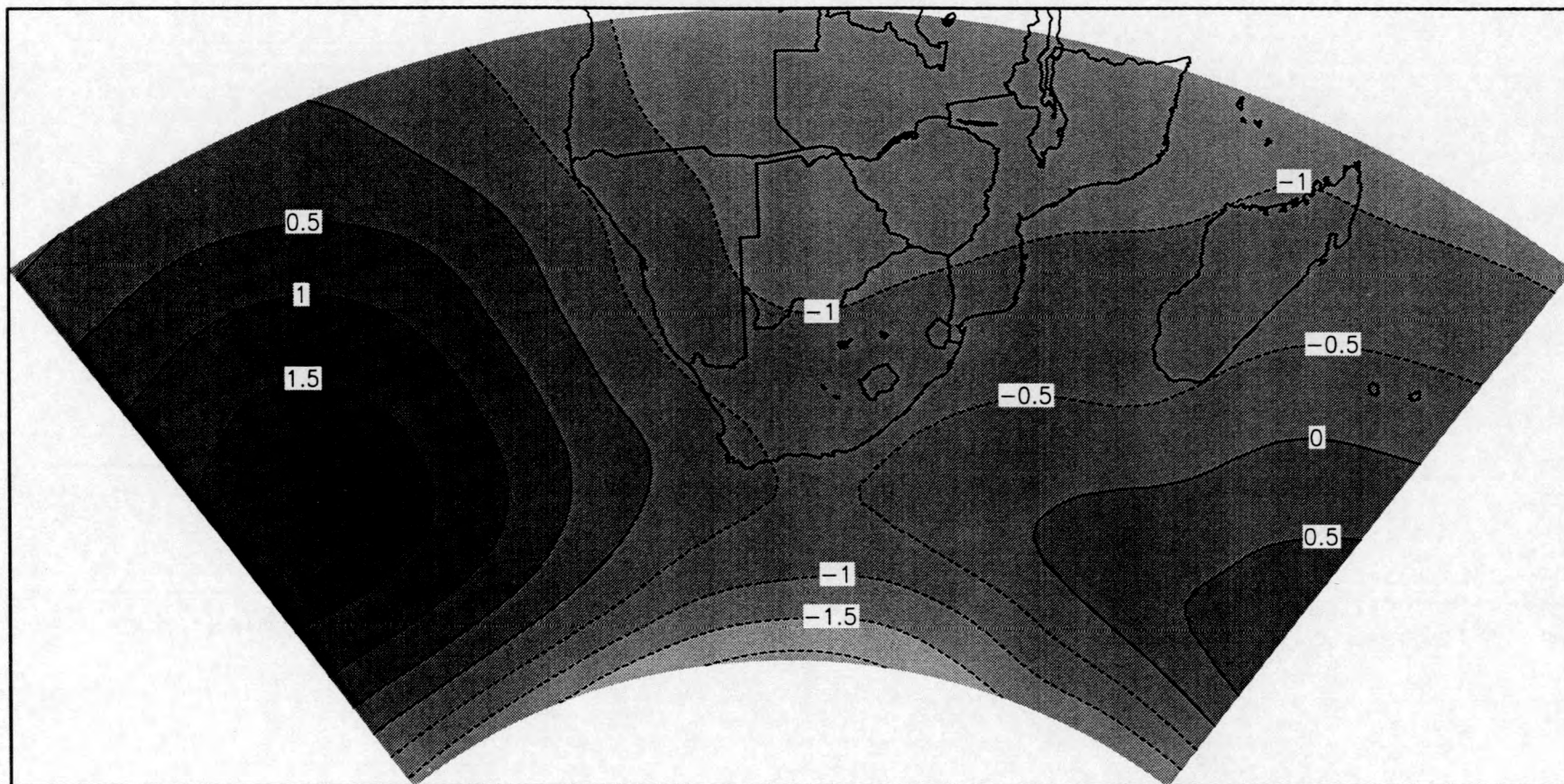


Figure 4.5 : A sample of the best matching input circulation pattern from January 1974 won by node (1,1)

An overall visual summary of the frequency of mappings to each node, differentiated by months, are presented in Figure 4.6 for the observed SLP data. There are 14 grid-boxes shown in Figure 4.6, 12 around the perimeter and two in the centre. The grid box at centre-right, labelled 'FRQ', shows the total distribution of wins across the node map, each grid box is referenced with co-ordinates corresponding to the two-dimensional node map and each node is coloured by the number of wins, darker shades representing a higher frequency wins. The grid box at centre-left, adjacent to this, labelled 'ERR' depicts, in relative shades, the average error associated with wins at each of the respective nodes. The figure is arranged with 12 maps reading clockwise around the border of the image (relating to each respective month, beginning in the top left corner with January). For these, the image shows the relative frequency and spatial distribution (in node space) of wins for a particular month across the node map. The location of clusters of hits may be used to identify the intra-annual variability and the dispersion of the clusters relates to the degree of inter-annual variability, where less dispersed clusters indicate less inter-annual variability.

To aid interpretation, Table 4.0 shows more detail associated with each of the frequency distribution plots and the total error and total frequency plots. The first three rows of values indicate, for a given month, a) the number of observations for that particular month; b) the maximum number of hits on any one node in the two-dimensional map for a given month and; c) the total number of nodes with at least one hit of which a maximum of 24 possible in the case of this SOM. The final row of Table 4.0 shows the range of values associated with the shading for the total error map 'ERR' and the accumulated frequency map 'FRQ'. The accumulated frequency map illustrates that no node was without at least

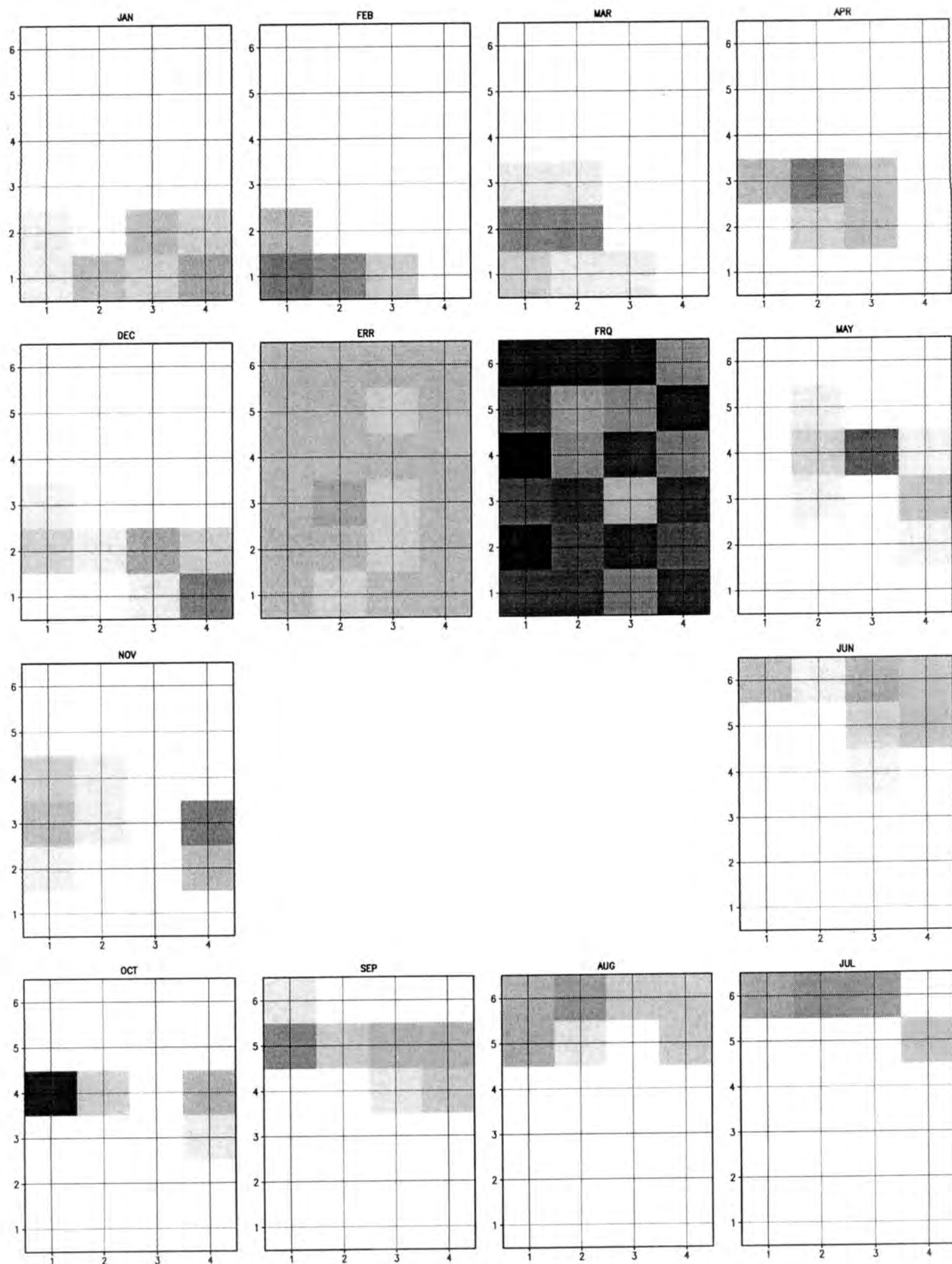


Figure 4.6 : Monthly frequencies and error map for Observed SLP circulation patterns.

three hits and none exceeded 13 hits. For example, the month of October (lower left-corner of Figure 4.6) the darkest shade indicates the cell with the maximum number of hits, referenced as node (1,4). From Table 4.0 the node with the maximum number of hits for that month is determined to have 11 hits, of which only 17 (one October in every year from 1972 to 1989) were possible. In this example, the high concentration of hits to such few nodes indicate a low degree of inter-annual variability since, in most years October is characterised by the same general synoptic type.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
# observations	16	16	16	16	17	17	17	17	17	17	17	17
Max. Hits/Node	4	6	5	6	8	4	5	4	5	11	6	6
Total Nodes Hit	7	4	7	5	7	7	4	7	7	4	7	7
Total	Range of Errors				2.394 to 3.528			Range of Frequencies			3 to 13	

Table 4.0 : Observed SLP Frequency map statistics.

The observation-to-node mappings are then used to calculate the centroid of the clusters for each month and plot them on a separate graph as shown in Figure 4.7, using the same co-ordinate system as described above. The graph shows the average annual cycle (intra-annual variability) expressed in the two-dimensional space of the SOM nodes, with summer months occupying the lower portion of the node space and winters occupying the upper portion of the node map. Some overlap of the transitional months of April/May and October/November is evident and to be expected. The meta-map may be referred to in order to find which synoptic pattern is associated with each specific node.

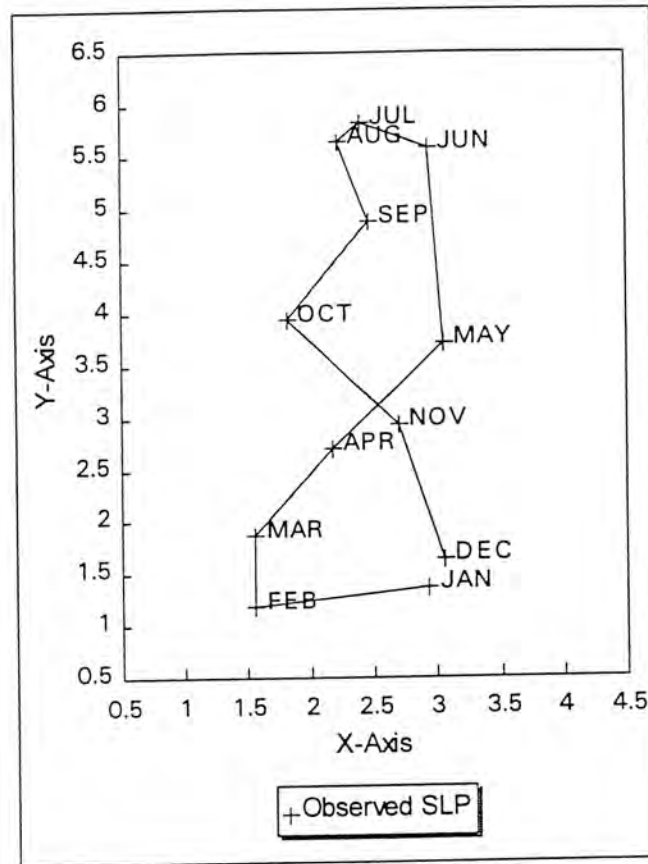


Figure 4.7 : Trajectory plots of centroids for each month from observed SLP.

GCM Control Simulation

Using the SOM trained with observed data, a calculation of frequency of mappings to each node for the GCM control simulation data are presented in Figure 4.8. Table 4.1 shows the physical quantities associated with frequencies and errors at each node. The month of September, for example, showed that one node (4,5) recorded all 5 possible hits, showing very little inter-annual variability while the month of April shows a high degree of inter-annual variability with four hits spread across the two-dimensional space. Generally errors are higher than those seen in the observed data set and a number of nodes (12 out of 24) did not win any map patterns at all.

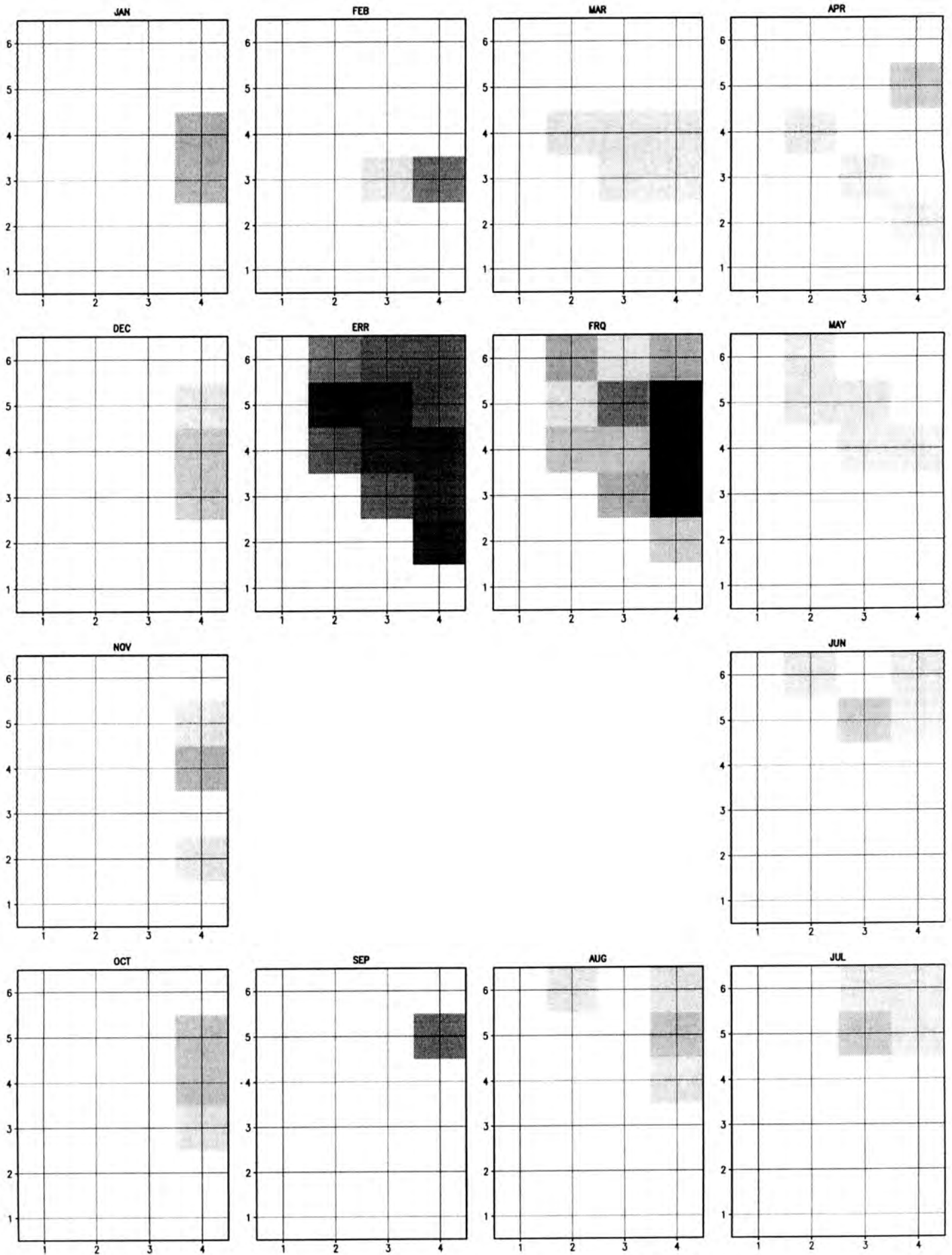


Figure 4.8 : Monthly frequencies and error map for control simulation SLP circulation patterns.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
# observations	6	6	5	5	5	5	5	5	5	5	5	5
Max. Hits/Node	3	5	1	2	1	2	2	2	5	2	3	2
Total Nodes Hit	2	2	5	4	5	4	4	4	1	3	3	3
Total	Range of Errors				5.337 to 8.145				Range of Frequencies		0 to 15	

Table 4.1 : GCM control run SLP frequency map statistics.

The centroid of the clusters for each month are then plotted on a separate graph as shown in Figure 4.9. The graph shows the average annual cycle, with summer months occupying the middle left portion of the node space and winters occupying the upper left of the node map.

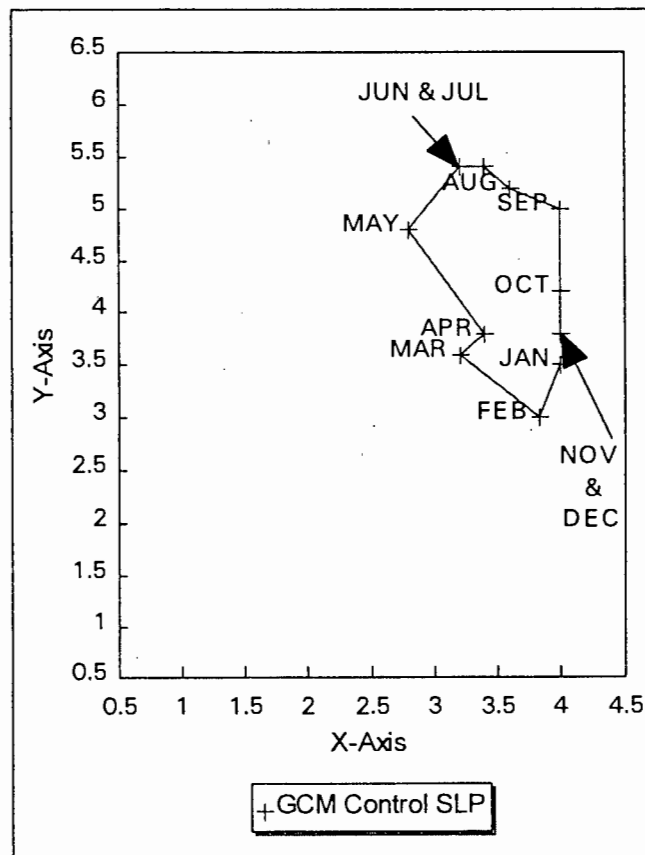


Figure 4.9 : Trajectory plots of centroids for each month from GENESIS GCM control SLP.

What is most characteristic of this data set are the lack of wins in the lower left, left and bottom of the node space as shown by the accumulated frequencies grid-box. Referring to

the meta-map (Figure 4.3), this area of the node space is represented by the more extremes of the sub-tropical trough (bottom) and the south Atlantic high pressure system (left). It is possible that this phenomenon may be explained by the weaker contrast of seasonal mean pressures that result from the way the SLP are calculated in a GCM. That is, while the observed data set is a function of point samples, the GCM data are a function of GCM grid cell average. The effect of grid-cell averaging is to reduce the extreme values that might have been found within a GCM grid-box. Lack of resemblance between the control run and the observed data may further be explained by the course resolution of GCM grid cells, which might have the effect of representing synoptic scale features at slightly different locations on the map. Furthermore, since the GENESIS GCM uses a simple mixed layer model to drive the atmosphere, it is possible that the full range of ocean forcing on the seasonal cycle has not been captured.

The annual cycle in the GCM is characterised by some similarities with the months of June, July and August of the observed data, while the remaining months do not correspond, possibly for reasons explained above. There is however a discernible seasonal cycle in the progression of the season with respect to synoptic types which may be seen in trajectory plots of Figure 4.9.

Thus, while the GCM simulation control run, on the face of it, fails to represent these extreme map patterns it has nevertheless captured the annual cycle within the area characterised by less intense pressure systems.

4.2 Doubled Carbon Dioxide Circulation

A new SOM, of similar structure to that used in the previous section, was then initialised and trained 20 times with the GENESIS GCM control simulation SLP ($1xCO_2$). The 20 trainings resulted in the a set of resultant quantisation errors ranging from 3.156 per sample to 3.236 per sample. However, the training that produced the lowest error did not yield the most satisfactory Sammon mapping, thus the sixth training was used, for which the average quantisation error was 3.167 per sample and the Sammon mapping shows an even spacing of nodes with respect to their characteristic weights (Figure 4.11). A time series of the average quantisation errors during training reveal the way in which the error is reduced through the training (Figure 4.10).

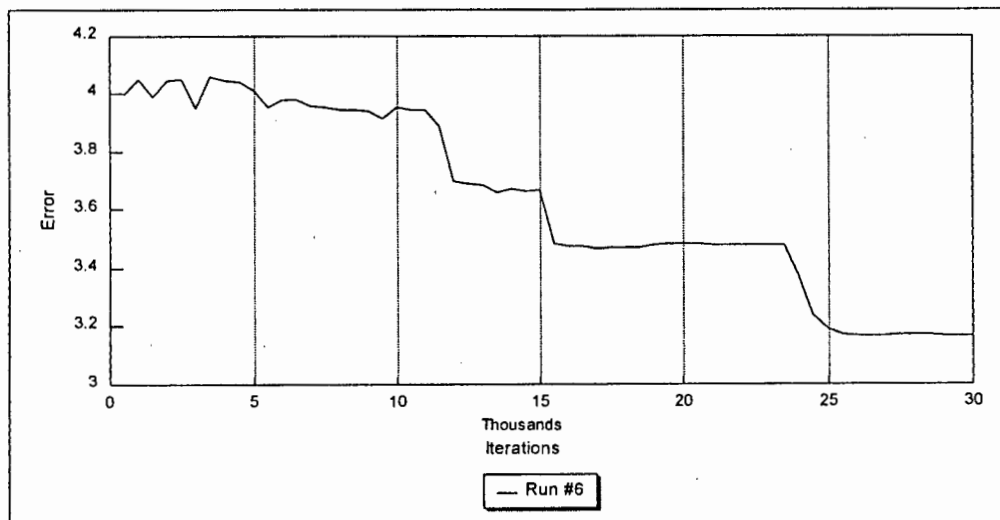


Figure 4.10 : Time series of errors calculated from the snapshots of the weights during training for the sixth run.

A quantisation error was then calculated for the GCM $2xCO_2$ simulation which gave an average quantisation error of 4.309 per sample, as opposed to the 3.167 of the training data. The relatively small difference in the overall error, when compared to the previous training using the observed circulation, indicates the similarity of circulation patterns of the $1xCO_2$

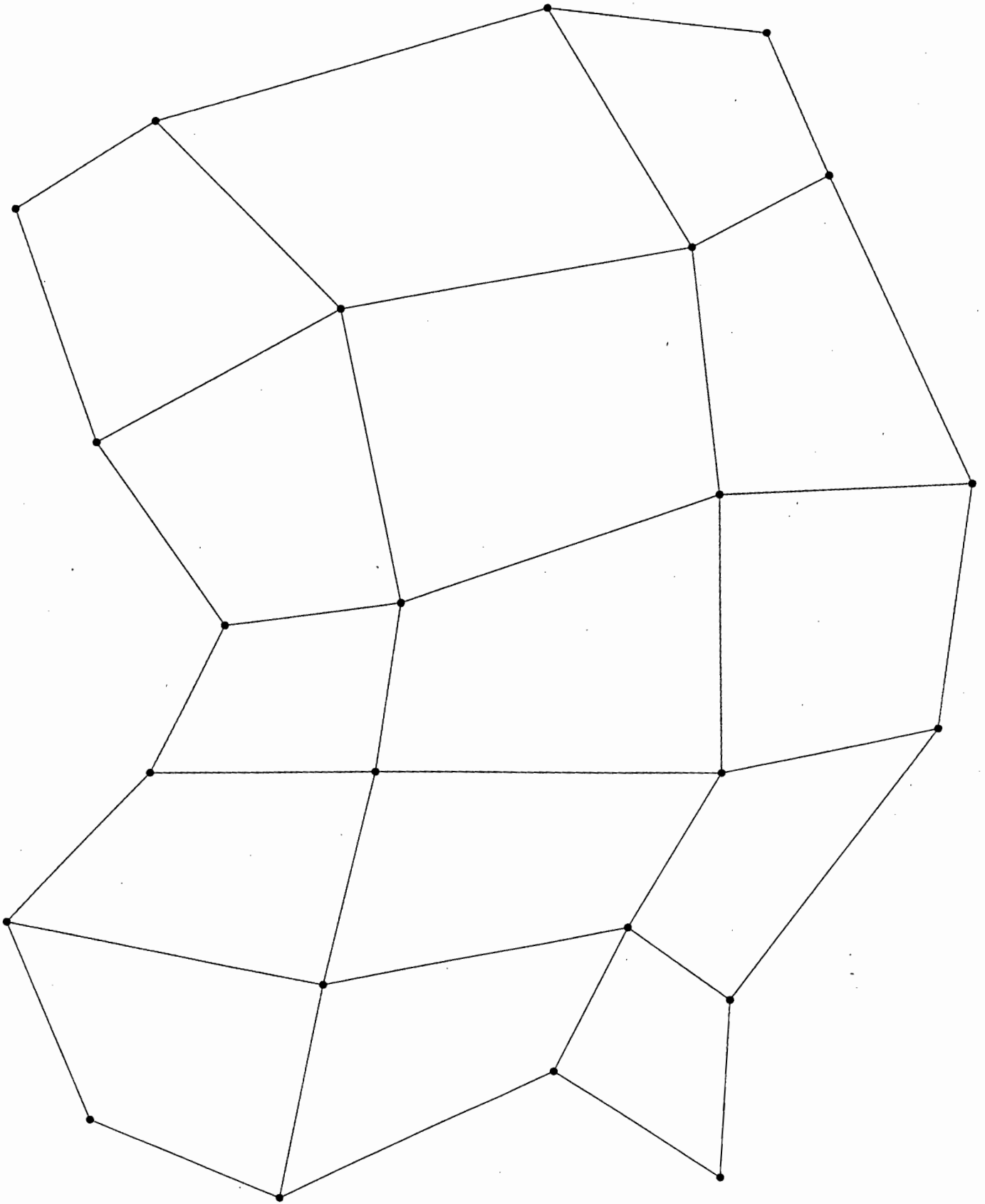


Figure 4.11 : Sammon mapping from the sixth training with GENESIS GCM control run simulation circulation data.

circulation to those in the 2xCO₂ data set. This suggests that the climate states of the 1xCO₂ and the 2xCO₂ data sets are more similar than the climate states of the observed and the 1xCO₂ circulation patterns, lending more support to the theory that difference in the observed and 1xCO₂ circulation patterns are a result of the characteristics of the GCMs particularly the grid averaging and coarse grid-scale resolution.

The weights of the SOM were then analysed by plotting the meta-map (Figure 4.12). As in the previous SOM training of the observed SLP data, each map shows varying interactions between the mid-latitude cyclones, the sub-tropical trough and the south Atlantic high and south Indian high pressure systems, with the difference being that the spatial configurations of the systems are slightly different due to the nature of the GCM with respect to grid-scale, parameterisation and grid areal averages, as discussed earlier (Section 2.2).

The set of observation-to-node mappings were then computed for both the 1xCO₂ and the 2xCO₂ data. A visual summary of the frequency of mappings to each node, differentiated by months, is presented in Figure 4.13. Table 4.2 shows the summary statistics including the number of possible samples, the maximum number of hits to a particular node and the total number of nodes hit on a month-by-month basis.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
# observations	6	6	5	5	5	5	5	5	5	5	5	5
Max. Hits/Node	2	2	2	2	2	2	3	2	2	3	2	1
Total Nodes Hit	4	5	4	4	4	4	3	4	4	2	3	5
Total	Range of Errors				2.223 to 5.646			Range of Frequencies			1 to 5	

Table 4.2 : GCM control run SLP frequency map statistics.

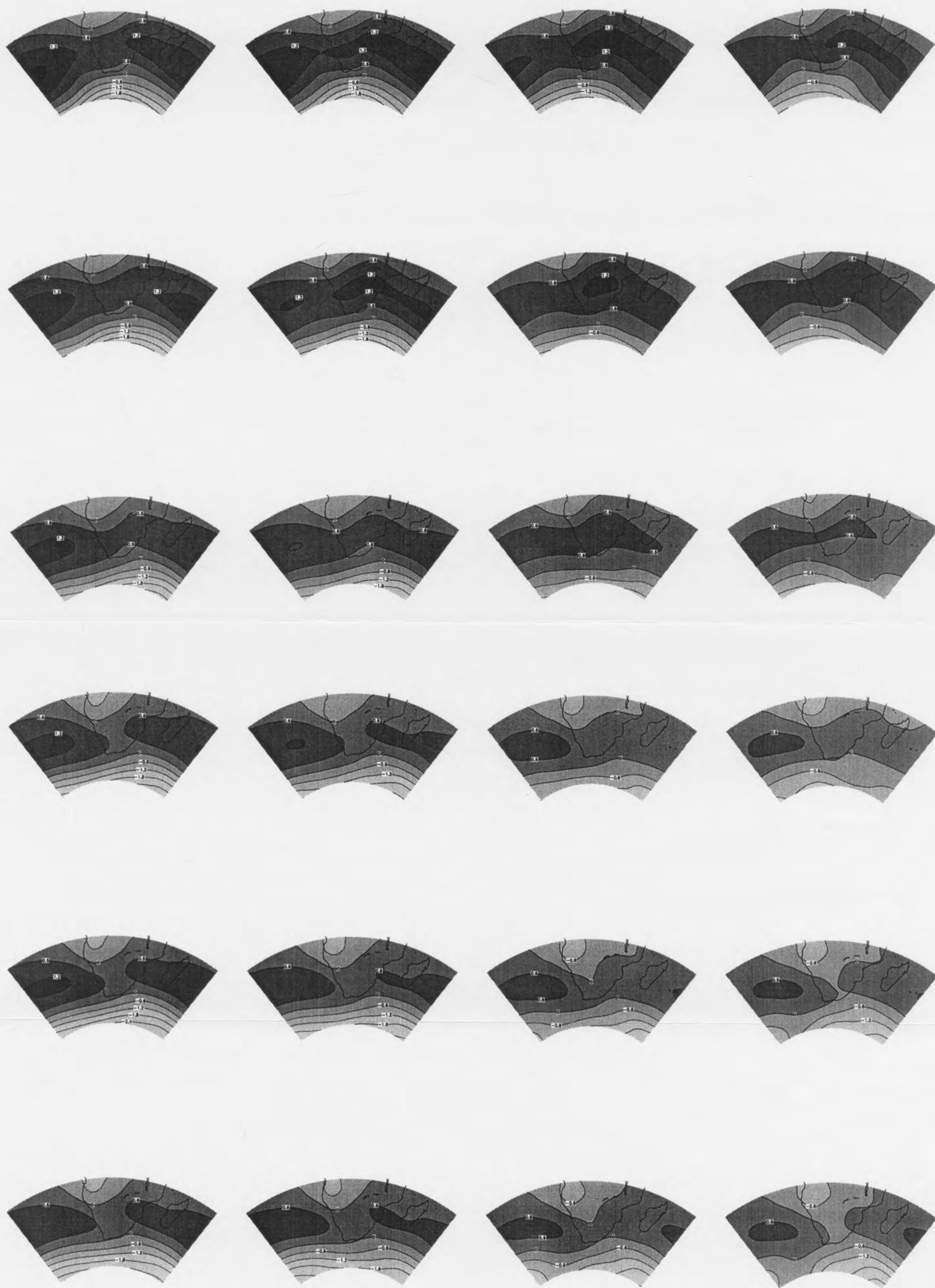


Figure 4.12 : Meta-map for GENESIS GCM control run SLP circulation.

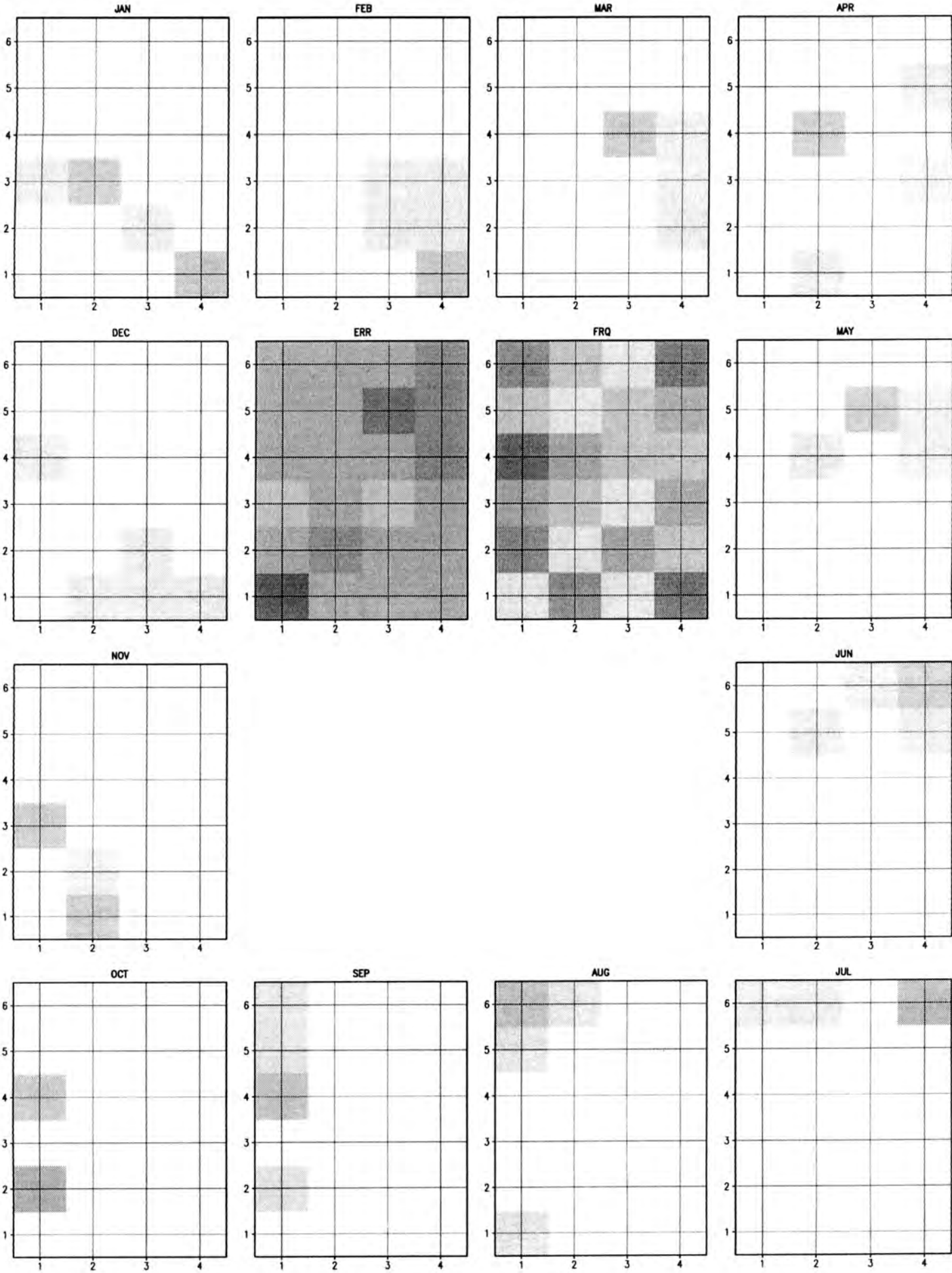


Figure 4 13 · Monthly frequencies and error man for control simulation SLP circulation patterns

The centroids of each cluster for each month are then plotted as shown in Figure 4.14 to illustrate the average annual cycle with respect to the node map. Due to the low number of monthly observations (five for all months except January and February which have six) in the GCM data, the trajectory is easily skewed when there is a large degree of inter-annual variability associated with a particular month, for example January and April have dispersed mappings when viewed on Figure 4.13. Despite this, however, the average annual cycle is clearly defined and may be verified by interpreting the cycle alongside the meta-map of Figure 4.12. In this way, the meta-map may be referred to in order to find which synoptic pattern is associated with a particular location on the node map.

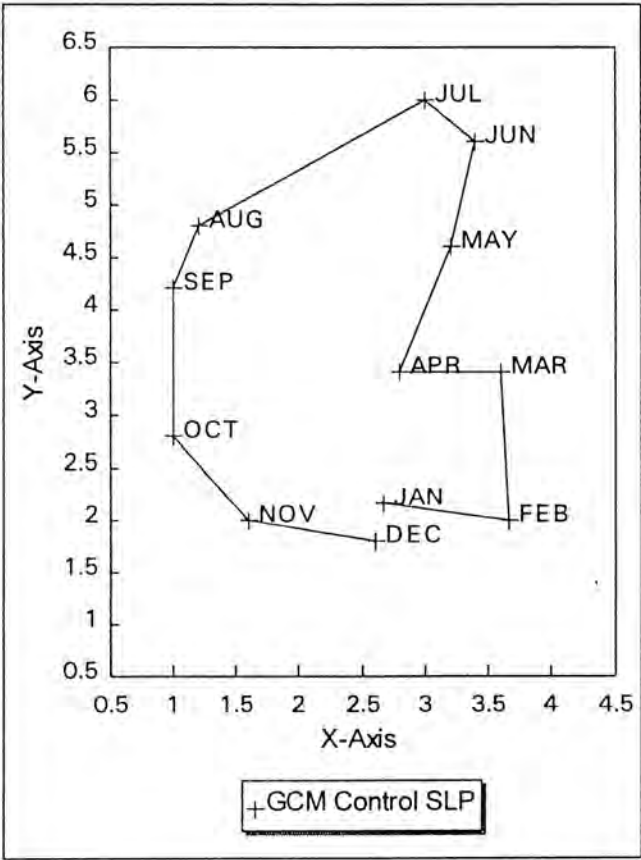


Figure 4.14 : Trajectory plots of centroids for each month of GENESIS GCM control SLP.

GCM Doubled CO₂

A calculation of frequency of mappings to each node are presented in Figure 4.15 for the GCM doubled carbon dioxide simulation (2xCO₂) data as projected through the SOM trained with the GCM control run simulation 1xCO₂ circulation patterns. Table 4.3 shows the summary statistics associated with frequencies for each month and the range of accumulated frequencies and average errors. The node located at (3,2), for example, did not win any of the circulation patterns found in the 2xCO₂ environment.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
# observations	6	6	5	5	5	5	5	5	5	5	5	5
Max. Hits/Node	3	2	2	3	3	2	2	1	1	2	2	2
Total Nodes Hit	4	4	4	3	3	4	4	5	5	4	4	4
Total	Range of Errors				6.125 to 7.222				Range of Frequencies			0 to 6

Table 4.3 : GCM doubled CO₂ run SLP frequency map statistics.

Figure 4.16 shows the trajectory of monthly centroids mapped to the node map. The graph shows a less well defined cyclical pattern to the seasonal cycle, when compared to that of the 1xCO₂ circulation data (Figure 4.14). In particular, while the annual cycle described by the trajectory of the centroids of each graph move anti-clockwise around the node map the exact path of the trajectory in the 2xCO₂ graph is clearly different and does not extend into the upper, upper-left and left side of the node map.

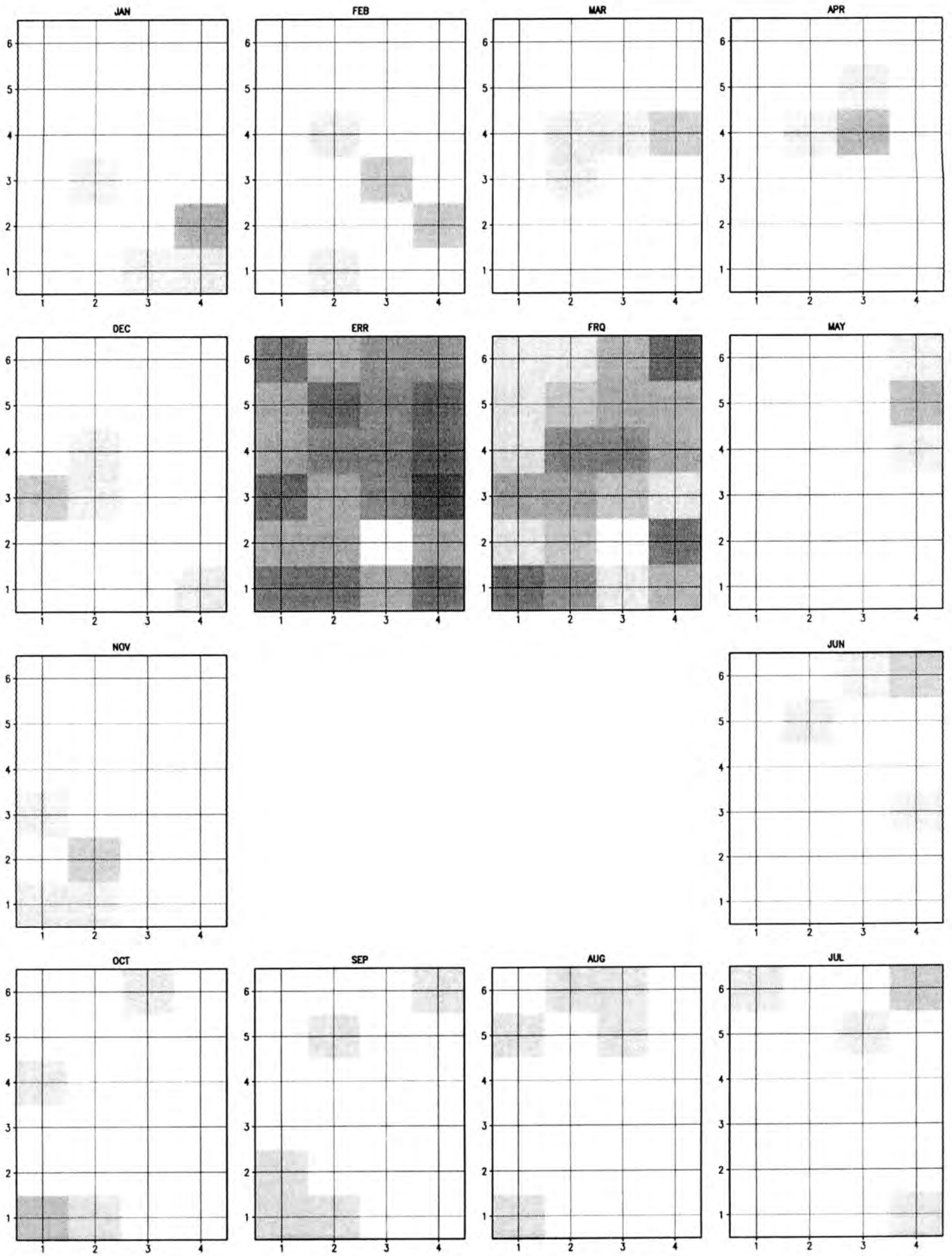


Figure 4.15 : Monthly frequencies and error map for doubled CO₂ simulation circulation patterns.

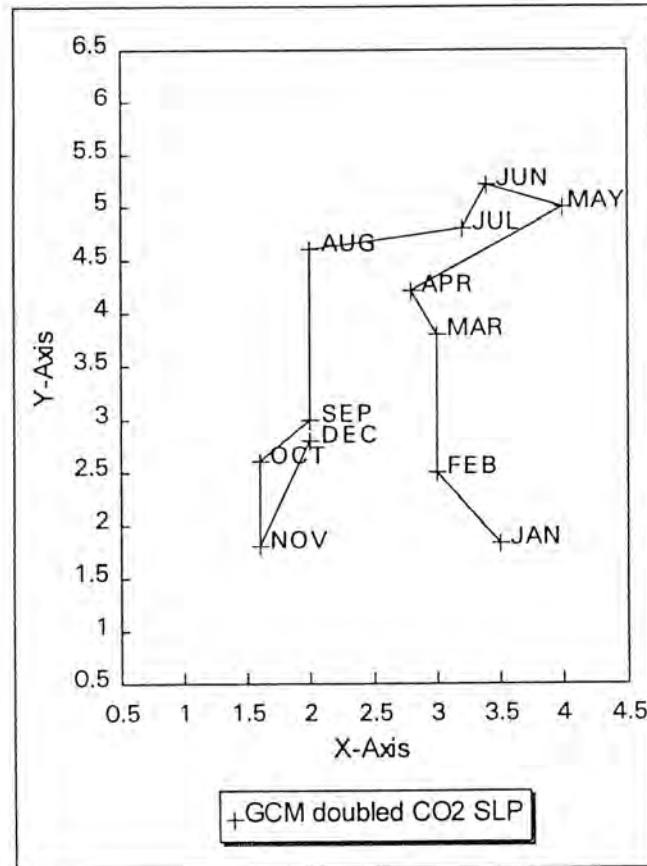


Figure 4.16 : Trajectory plots of centroids for each month of GENESIS GCM doubled CO₂ SLP.

When comparing trajectories of the average annual cycle of the 1xCO₂ (Figure 4.14) and 2xCO₂ (Figure 4.16), there are some apparent differences. The 1xCO₂ trajectory shows a clear cyclical trajectory around the SOM map, with the exception of April, whose centroid is perturbed by the dispersed mappings associated with a high degree of inter-annual variability, as can be seen in Figure 4.13. In comparison, the 2xCO₂ data shows a trajectory with the same general form except that centroids here are affected by the highly dispersed mappings except for those in March, April, May and November (Figure 4.15). The high sensitivity of centroids to dispersed mappings is even more pronounced in the GCM data than the observed data as there are only five or six observations per month compared to the 16 or 17 observations per month.

Nonetheless, disregarding the effect of small sample size and inter-annual variability, some interpretations may be made about changes in each of the four groups of December, January and February (DJF); March April and May (MAM); June, July and August, (JJA) and September, October and November (SON). The $2xCO_2$ DJF season is characterised by a break in the seasonal pattern when compared to the $1xCO_2$ patterns with December being dominated more by stronger high pressure systems (see meta-map, Figure 4.12) and $2xCO_2$ January tending more towards deeper sub-tropical trough systems that feed moisture into the continental interior with little difference between the $1xCO_2$ and $2xCO_2$ patterns in February. MAM is characterised by a general movement of centroids towards the top of the map from the $1xCO_2$ map to the $2xCO_2$ map. Referring to the meta-map this translates to more intense high pressure systems over the sub-continent, which is generally indicative of elevated inversions and continental drying. The JJA season is characterised by a shift in the centroids away from the top of the SOM map in the $2xCO_2$ simulation when compared with the $1xCO_2$ simulation, denoting a trend towards less pronounced westerly systems which take the form of mid-latitude cyclones. Finally the SON season, as the precursor to summer, has shown a general weakening of both high and low pressure systems in the $2xCO_2$ simulation in September and October with very little difference between the $1xCO_2$ and $2xCO_2$ simulations in November.

This type of general analysis of the changes in seasonality between the $1xCO_2$ circulation patterns and the $2xCO_2$ circulation patterns is a useful exercise for acquiring a 'first-cut' view of the differences in the data sets which can then further aid understanding of subsequent analyses.

4.3 Discussion and Summary

Clearly, from the SOM analysis of the observed and GCM control run data, distinct differences in the seasonal cycle are apparent. The control run circulation patterns are generally less defined with regard to the observed data with respect to strong differences in relative intensities of high and low pressure systems. Most GCM control simulation patterns have mapped to nodes characterised by less intense circulation patterns. However, the seasonal cycle is still evident in the general circulation, suggesting that the overall general circulation is being modelled correctly, although lacking in well defined systems. The lack of definition could perhaps be an indication of the lack of ocean forcing on the atmosphere as simulated by the simple mixed layer model of the GENESIS v1.02 GCM, and as noted in Chapter 2, oceans are important in southern African climate and are not modelled well. With the characteristic of the model in mind, the focus is then turned towards the differences between the control run and the doubled CO₂ run.

Clearly the lack of a long record of data for both the 1xCO₂ and 2xCO₂ data sets has hampered interpretation of changes in inter-annual variability. However, the application of the SOM in this context has shown that the evaluation of the intra-annual changes in seasons may be made.

Finally, the SOM has illustrated a useful power to generalise synoptic patterns into a form of categories that may be expressed in two-dimensions and from which trajectories of the type shown may be plotted for further interpretation.

CHAPTER FIVE

Conclusion

5.0 Discussion

Based on the procedure and results presented above, the SOM is shown to be yet another useful computer-based synoptic climatology technique. In the light of the advent of computer assisted techniques in synoptic climatology, the SOM technique has expanded the group of methods by which synoptic climatology may be approached, although such techniques still remain ultimately subjective. The use of a SOM as a synoptic climatology technique is no exception, and a number of decisions made by the investigator could significantly impact upon the final interpretation. These decisions are primarily related to the training of the SOM with respect to topology, dimensions of target domain, training type and length and preparation of data.

The main purpose of synoptic climatology is to relate the circulation to the surface environment either by a circulation-to-environment approach or by an environment-to-circulation approach for which many methods are available. In the example of the application of the SOM to a climate change problem the classification of a circulation-to-environment implementation of the SOM was shown. The synoptic map patterns were classified into types, each co-related to one another. The SOM technique has illustrated its ability to capture modes of variability in the map patterns and express these synoptic types

across a two-dimensional continuum, in such a way that the seasonality of the synoptic types may be seen clearly. Furthermore, analysis of the observation-to-node mappings, shows potential to yield a more detailed mechanism for determining the nature of inter-/intra-annual variability.

In the context of global climate change, GCM output is likely to be one of the primary sources of data for the development of scenarios. The SOM technique is shown to highlight key areas of difference and similarity within a circulation data set, providing information needed to validate GCM data. The analysis is primarily hampered by the coarse resolution of the GCM data and the subsequent effects of areal averaging. However, with ever improving resolution capabilities of the GCM, these problems are likely to become less significant.

5.1 Conclusion

In conclusion to this study, a methodology using the Kohonen Self-Organising Map has been introduced which makes use of techniques of Artificial Neural Networks (ANNs) to classify input data on a non-linear basis. The methodology requires samples of circulation data for training, after which any sample, within the bounds of the training data may be presented to the SOM which will classify the sample and associate a measure of success with the classification.

Due to the properties of the SOM, particularly the non-linear classification and the continuous nature of the categories, the methodology assists in circumventing some of the difficulties of the more traditional synoptic climatological techniques. Such traditional techniques are either difficult to interpret due to their mathematical compactness, or derive discontinuous categories of synoptic typing. The SOM methodology does have the common property in which the technique is subjective in its process and as such relies on the investigator to have a working knowledge of the data being analysed and the possible variations available when using the technique. The capabilities of the SOM make the technique suitable for use with other computer-assisted synoptic climatology techniques.

When applied to an example of investigating the inter- and intra-annual variability of the atmospheric circulation over southern Africa, both recent past and future, the SOM methodology proved to be useful for visualising and determining the nature of variability of synoptic patterns on these time scales. The SOM methodology is particularly useful for comparing data sets on the basis of this variability and shows potential for use in global climate change studies, and in particular, for use with General Circulation Models (GCMs). The GCM, while restricted by the ability to model sub-grid-scale processes, is adequately capable of modelling large scale circulation, which in turn is the key to driving surface conditions through synoptic controls.

With future global climate change studies likely to be making use of GCMs as the principal tool, the SOM methodology has been shown to be a useful technique for validating the GCM on the basis of synoptic circulation, and for comparing control and experimental

simulations to aid the interpretation and formulation of global climate change scenarios.

As computational resources improve, the type of application, shown above, could be extended to classification at shorter time scales and improved spatial resolutions. Overall the procedure appears to be robust for a wide range of possible synoptic climatological uses.

References

- Bardossy, A., Duckstein, L. and Bogardi, I., 1995:** Fuzzy Rule-Based classification of atmospheric circulation patterns, *Int. J. of Climatology*, **15**, 1087-1097
- Barry, R.G. and Perry, A.H., 1973:** *Synoptic climatology: Methods and applications*, Methuen & C, London.
- Brankovic C., Palmer T.N., Ferranti L., 1994:** Predictability of seasonal atmospheric variations, *Journal of Climate*, **7**, 217-237.
- 'Bayo Omotosho, J., 1992:** Long-range prediction of the onset and the end of the rainy season in the west African Sahel, *Int. J. of Climatology*, **12**, 369-382
- Buell, C.E., 1979:** On the physical interpretation of empirical orthogonal functions, *Sixth Conference on Probability and Statistics in Atmospheric Science*, American Meteorological Society, Boston, 112-117.
- Clothiaux, E.E. and Bachmann, C.M., 1994:** Neural Networks and their applications, In Hewitson, B.C. and Crane R.G. (eds), *Neural Nets: Applications in Geography*, Kluwer Academic Publishers, Dordrecht.
- Currie, R.G., 1993:** Luni-solar 18.6 and 10-11-year solar cycle signals in South African rainfall, *Int. J. of Climatology*, **13**, 237-256
- D'Arbeton P.C. and Lindesay J.A., 1993:** Water vapour transport over southern Africa during wet and dry early and late summer months, *Int. J. of Climatology*, **13**, 151-170
- Eagleman, J.R., 1976:** *The visualization of climate*, Lexington Books, Toronto.

- Gates, W.L.**, 1992: AMIP: The Atmospheric Model Intercomparison Project, *Bulletin American Meteorological Society*, **73**, 12, 1962-1970
- Giorgi, F., Shields Brodeur C. and Bates G.T.**, 1994: Regional climate change scenarios over the United States produced with a nested regional climate model, *Journal of Climate*, **7**, 375-399.
- Grotch, S.L. and MacCracken, M.C.**, 1991: The use of General Circulation models to predict regional climate change, *Journal of Climate*, **4**, 286-303.
- Hertz, J.A, Krogh, A.S. and Palmer, R.G.**, 1991: *Introduction to the theory of neural computation*, Addison-Wesley, Reading (Massachusetts).
- Hewitson, B.C.**, 1994: Regional Climates in the GISS General Circulation Model: Air Temperature, *Journal of Climate*, **7**, pp 283-303.
- Hewitson B.C. and Crane R.G.**, 1992a: Large-scale atmospheric controls on local precipitation in tropical Mexico, *Geophysical Research Letters*, **19**, 18, 1835-1838.
- Hewitson B.C. and Crane R.G.**, 1992b: Regional-scale climate prediction from the GISS GCM, *Palaeogeography, Palaeoclimatology, Palaeoecology (Global planetary change section)*, **97**, 249-267.
- Hewitson, B.C. and Crane R.G. (eds)** 1994: *Neural Nets: Applications in Geography*, Kluwer Academic Publishers, Dordrecht.
- Hewitson, B.C. and Crane, R.G.**, 1996: Climate downscaling: techniques and application, Climate Research, in press.

- Hewitson, B.C. and Main J.P.L.**, 1996: Regional Climate Change scenarios for precipitation and temperature from GCMs, Interim Report, Water Research Commission, Pretoria, South Africa.
- Houghton, J.T., Callender, B.A. and Varney, S.K.**, 1992: *Climate Change 1992, The supplementary report to the IPCC scientific assessment*, Cambridge University Press, Cambridge.
- Hurrell J.W. and van Loon H.**, 1994: A modulation of the atmospheric annual cycle in the Southern Hemisphere, *Tellus*, **46A**, 325-338.
- Hulme, M.**, 1992: Rainfall changes in Africa: 1931-1960 to 1961-1990, *Int. J. of Climatology*, **12**, 685-699.
- Jenne, R.L.**, 1975: Data sets for Meteorological Research, NCAR-TN/1A-111, National Center for Atmospheric Research, Boulder, Colorado.
- Jolliffe, I.T. and Sarria-Dodd, D.E.**, 1994: Early detection of the start of the wet season in tropical climates, *Int. J. of Climatology*, **14**, 71-76.
- Joubert, A.M. and Hewitson, B.C.**, 1996: Simulating present and future climates of Southern Africa using general circulation models, *Progress in Physical Geography*, in press.
- Kalkstein, L.S.**, 1979: A synoptic climatological approach for environmental analysis, *Proc. Middle States Div.: Association of American Geographers*, **13**, 68-75.
- Kalkstein, L.S., Tan, G. and Skindlov, J.A.**, 1987: An evaluation of objective clustering procedures for use in synoptic climatological classification, *J. Climate and Applied Meteorology*, **26**, 717-730.

- Kalkstein, L.S. and Corrigan, P.**, 1986: A synoptic climatological approach for environmental analysis: Assessment of sulphur dioxide concentrations, *Annals of the Association of American Geographers*, **76**, 381-395.
- Kerr, R.A.**, 1995: Ice, quakes, and a wobble shake San Francisco in *Science*, **267**, 27-28.
- Key, J. and Crane, R.G.**, 1986: A comparison of synoptic classification schemes based on 'objective' procedures, *Journal of Climatology*, **6**, 375-388.
- Kirchhofer, W.**, 1973: Classification of European 500 mb patterns, *Arbeitsbericht der Schweizerischen Meteorologischen Zentralanstalt Nr. 43*, Geneva.
- Kohonen, T., Hynninen J., Kangas J. and Laaksonen J.**, 1996: SOM_PAK: The Self Organising Map Program Package. Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, Finland.
- Kohonen, T.**, 1990: The Self-Organizing Map, *Proceedings of the IEEE*, **78**, 9, 1464-1480.
- Kohonen, T.**, 1984: *Self-Organization and Associative Memory*, Springer Verlag, Berlin.
- Lamb, H.H.**, 1972: British Isles weather types and a register of the daily sequence of circulation patterns, 1861-1971, *Geophysical Memoirs*, **116**, London
- Le Marshall, J.F., Kelly, G.A.M. and Karoly, D.J.**, 1985: An atmospheric climatology of the southern hemisphere based on ten years of daily numerical analysis (1792-1989): I Overview, *Australian Meteorological Magazine*, **33**, 65-85.
- Lindesay, J.A.**, 1988: Southern Africa Rainfall, the southern oscillation and the southern hemisphere semi-annual cycle, *Journal of Climatology*, **8**, 17-30.

- Lindesay J.A. and Vogel, C.H.**, 1990: Historical evidence for southern oscillation-southern African relationships, *Int. J. of Climatology*, **10**, 679-689
- Lund, I.A.**, 1963: Map-pattern classification by statistical methods, *Journal of Applied Meteorology*, **2**, 56-65.
- Mason, S.J.**, 1990: Temporal variability of sea surface temperatures around southern Africa: a possible forcing mechanism for the 18-year rainfall oscillation?, *South African J. of Science*, **86**, 243-252
- Mason, S.J. and Lindesay, J.A.**, 1993: A note on the modulations of Southern Oscillation-Southern African rainfall associations with the Quasi-biennial Oscillation, *Journal of Geophysical Research*, **98**, D5, pp 8847-8850.
- Mason S.J. and Tyson, P.D.**, 1992: The modulation of sea surface temperature and rainfall associations over southern Africa with solar activity and the quasi biennial oscillation, *Journal of Geophysical Research*, **97**, D5, pp 5847-5856.
- Matarira, C.H. and Jury, M.R.**, 1992: Contrasting Meteorological structure of intra-seasonal wet and dry spells in Zimbabwe, *Int. J. of Climatology*, **12**, 165-176
- Matarira, C.H.**, 1990: Drought over Zimbabwe in a regional and global context, *Int. J. of Climatology*, **10**, pp 609-625
- Mearns, L.O., Giorgi, F., McDaniel, L. and Shields, C.**, 1995a: Analysis of daily variability of precipitation in a nested regional climate model: comparison with observations and doubled CO₂ results, *Global and Planetary Change*, **10**, 55-78

- Mearns, L.O., Giorgi, F., McDaniel, L. and Shields, C., 1995b:** Analysis of variability and diurnal range of daily temperature in a nested regional climate model: comparison with observations and doubled CO₂ results, *Climate Dynamics*, **11**, pp 193-209.
- Mitchell, J.F.B., Johns, T.C., Gregory, J.M. and Tett, S., 1995:** Climate response to increasing levels of greenhouse gases and sulphate aerosols. *Nature*, **376**, 501-504.
- Muller, R.A., 1977:** A synoptic climatology for environmental baseline analysis: New Orleans, *Journal of Applied Meteorology*, **16**, 20-33.
- Murphy, J.M. and Mitchell, J.F.B., 1995:** Transient response of the Hadley Centre coupled ocean-atmosphere model to increasing carbon dioxide. Part II: Spatial and temporal structure of response, *Journal of Climate*, **8**, 57-80.
- Nicholson, S.E., 1986:** The nature of rainfall variability in Africa south of the equator, *Journal of Climatology*, **6**, 515-530.
- Pollard, D. and Thompson, S.L., 1995:** A climate model (GENESIS) with a land-surface transfer scheme (LSX). Part I: Present day climate, *Journal of Climate*, **8**, 732-761.
- Peixoto, J.P. and Oort, A.H., 1992:** *Physics of Climate*, American Institute of Physics, New York.
- Richman, M.B., 1986:** Rotation of Principal Components, *Journal of Climatology*, **6**, 293-335.
- Sammon, J.W. Jr., 1969:** A non-linear mapping for data structure analysis, *IEEE Transactions on Computers*, **C-18**(5), 401-409, May 1969.

- Santer, B.D. and Wigley, T.M.L.**, 1990: Regional validation of means, variances, and spatial patterns in general circulation model control runs, *J. of Geophysical Research*, **95**, D1, 829-850.
- Taljaard, J. J.**, 1986: Change of rainfall distribution and circulation patterns over southern Africa in summer, *Journal of Climatology*, **6**, 579-592.
- Theron G.F. and Harrison M.S.J.**, 1991: Thermodynamic properties of the mean circulation around southern Africa, *Theoretical and Applied Climatology*, **43**, 161-174.
- Tyson, D.A. and Hewitson, B.C.**, 1996: Mid latitude cyclones south of Africa in the GENESIS GCM, *Int. J. of Climatology*, in press.
- Tyson, P.D.**, 1993: Recent developments in the modelling of the future climate of southern Africa, *South African Journal of Science*, **89**, 494-505.
- Tyson, P.D.**, 1984: The atmospheric modulation of extended wet and dry spells over South Africa, *Journal of Climatology*, **1**, 621-635.
- Tyson, P.D.**, 1981: Atmospheric circulation variations and the occurrence of extended wet and dry spells over southern Africa, *Journal of Climatology*, **1**, 115-130.
- von Storch, H.**, 1993: Downscaling of global climate change estimates to regional scales: an application to Iberian rainfall in wintertime, *Journal of Climate*, **6**, 1161-1171.
- Wasserman, P.D.**, 1989: *Neural Computing Theory and Practice*, Von Nostrand Reinhold, New York.

- Walker, N.D.**, 1990: Links between South African summer rainfall and temperature variability of the Agulhas and Benguela current systems, *Journal of Geophysical Research*, **95**, C3, pp 3297-3319.
- Wigley, T.M.L., Jones, P.D., Briffa, K.R., Smith G.**, 1990: Obtaining sub-grid-scale information from coarse-resolution general circulation model output, *Journal of Geophysical Research*, **95**, D2, pp 1943-1953
- Whetton, P., Pittock, A.B., Labraga, J.C., Mullan, A.B. and Joubert, A.**, 1996: Southern Hemisphere Climate: Comparing models with reality, in Henderson-Sellers, A. and Giambelluca, TW, (eds), *Climate Change: Developing Southern Hemisphere perspectives*, Chichester, John Wiley and Sons.
- Willmot, C.J., Clinton, M.R., Philpot, W.D.**, 1985. Small-scale climate maps: A sensitivity analysis of some common assumptions associated with grid point interpolation and contouring, *The American Cartographer*, **12** , 1, 5-16.
- Willmot, C.J.**, 1982: Some comments on the evaluation of model performance, *Bulletin American Meteorological Society*, **63**, 11, 1309-1313.
- Willmot, C.J.**, 1981: On the validation of models, *Physical Geography*, **2**, 2, 184-194.
- Yarnal, B.**, 1993: *Synoptic Climatology in environmental analysis: A primer*, Belhaven Press, London.